2017

Identifying Predictors of Weight Loss and Drop-Out Using Joint Modeling

Valerie Bares
South Dakota State University

IDENTIFYING PREDICTORS OF WEIGHT LOSS AND DROP-OUT USING JOINT MODELING

BY

VALERIE BARES

A dissertation submitted in partial fulfillment of the requirements for the

Doctor of Philosophy

Major in Computational Science and Statistics

South Dakota State University

2017

IDENTIFYING PREDICTORS OF WEIGHT LOSS AND DROP-OUT USING JOINT MODELING

VALERIE BARES

This dissertation is approved as a credible and independent investigation by a candidate for the Doctor of Philosophy in Computational Science and Statistics degree and is acceptable for meeting the dissertation requirements for this degree. Acceptance of this does not imply that the conclusions reached by the candidate are necessarily the conclusions of the major department.

Xijin Ge, Ph.D. Date
Dissertation Advisor

Kurt Cogswell, Ph.D. Date
Head, Department of Mathematics & Statistics

Dean, Graduate School Date

ACKNOWLEDGEMENTS

I want to thank my advisor, Dr. Xijin Ge, for encouraging me through each step of my research but also giving me the freedom to make my own decisions. Each of my committee members helped me in their own way through this process. Thank you to Drs. Gary Hatfield, Howard Wey, Bonny Specker, and Febina Mathew for meeting with me and giving me suggestions along the way. I would also like to thank other members of the research group, especially Dongmin Jung.

Thank you to the South Dakota State University Mathematics and Statistics Department. Specifically, Dr. Kurt Cogswell for his support and Dr. Thomas Brandenburger for helping me through countless situations.

Several Sanford employees along the way were invaluable to me. Thank you to Dr. Paul Thompson for being generous with his time and knowledge. Profile by Sanford employees Dr. Stephen Herrmann, Chris Clark, and Natalie Papini were especially helpful in answering questions and engaging my interest in this topic.

Finally, I need to thank my family and friends for the constant support and encouragement. Many people helped me through numerous transitions throughout the last few years. I could not have done this without those people, past and present.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABBREVIATIONS

AIC	Akaike Information Criterion
AUC	Area Under the Curve
BMI	Body Mass Index
CDC	Centers for Disease Control and Prevention
CDF	Cumulative Distribution Function
CSV	Comma-Separated Values
FFM	Fat-Free Mass
FM	Fat Mass
PDF	Probability Density Function
RMSE	Root Mean Square Error
ROC	Receiver Operating Characteristic
SQL	Structured Query Language
VPN	Virtual Private Network
WHO	World Health Organization
WHR	Waist-to-Hip Ratio

ABSTRACT

IDENTIFYING PREDICTORS OF WEIGHT LOSS AND DROP-OUT USING JOINT MODELING

VALERIE BARES

2017

Profile by Sanford is a membership based weight loss program that helps its members make lifestyle changes with diet, exercise, and one-on-one interactions with a weight loss coach. Discovery of characteristics and behaviors influencing weight loss will benefit current and future members of Profile. This research utilizes massive data from Profile by Sanford to analyze member behavior. Fourteen data sets are evaluated, some containing millions of observations. All data is combined into one comprehensive table of 33,487 members. Members of Profile by Sanford are 77% female and two-thirds of all members start the program classified as obese.

Attending meetings with a weight loss coach decreases rapidly over time for Profile members but a higher frequency of meetings is found to have a positive association with weight loss. Increasing a member's coach meeting attendance to one more meeting a month results in 2.5 percentage points more weight loss for Profile members who weigh themselves consistently each month for the first 12 months in the program. The same group of Profile members experience 2.3 percentage points less weight loss if taking antidepressants after controlling for sex and starting BMI.

A mixed model generates weight loss predictions. An additional attendance of a coach meeting is associated with 0.13 percentage points more monthly weight loss. With

one more weight recording members lose 0.02 percentage points more per month. A unit increase in starting BMI is associated with an increase of 0.03 percentage points more weight loss.

By month 6 more than half of members have dropped out of Profile and 80% have dropped out by month 12. The probability of dropping out of the program is produced by a joint model. Higher age, married members, and females are associated with a lower risk of dropping out of Profile. The joint model suggests that the risk of dropping out of the weight loss program increases by 140% with each percentage point increase in monthly weight gain. Application of the statistical models can allow coaches to interact proactively with members based on their likelihood of dropping out of the program.

1 BACKGROUND

1.1 OBESITY

Obesity is a prevalent problem in America, where more than one-third of adults are considered obese [1]. Obesity rates remain high and the presence of programs to support weight loss have increased [2,3]. According to the Centers for Disease Control and Prevention (CDC) and the World Health Organization (WHO), adults are classified based on their height and weight, referred to as Body Mass Index (BMI) [4,5]. BMI is calculated by dividing a person's weight in kilograms by the square of their height in meters [4,5]. If a person's BMI is less than 18.5 they are considered underweight; between 18.5 and 25 is normal; BMI of 25 but less than 30 is considered overweight, and a BMI of 30 or higher is obese [4]. This index does not consider sex or age and does not consider actual body fat mass (FM) or fat-free mass (FFM). BMI was developed when its correlation with body fat mass was discovered [6]. Even with a significant correlation with FM, BMI can potentially overestimate body fat in individuals that are muscular and underestimate body fat in older individuals that have lost muscle [7].

Obesity classifications also exists by measuring a person's waist-to-hip ratio (WHR). The WHR is calculated by dividing waist circumference by hip circumference [6]. This ratio represents fat distribution better than BMI [6]. According to WHO, a WHR larger than 0.90 for men and 0.85 for women is classified as obese [8]. WHO specifies methods of measuring a person's waist and hip circumference and these measurements could vary if this protocol is not carefully followed [8]. WHR obesity classification differs by sex since men and women have different body compositions.

Men have higher total lean mass and bone mineral mass and lower fat mass than women [8]. Body composition differences also exist among age and ethnicity groups [8].

Obesity is associated with adverse physical and health problems [6]. High BMI and WHR are risk factors for diabetes and cardiovascular disease such as hypertension [8]. Financial costs also increase for obese people. The average annual medical costs for an obese individual is $1429 more than an individual who is classified as having a normal BMI [9].

1.2 WEIGHT LOSS

Diet restriction is essential to weight loss. Sacks *et al.* followed four groups of people on diets with different composition of fat, protein, and carbohydrates, but found no difference in body weight after two years [10]. Some diets encourage low carbohydrate intake. Meta-analysis by Clifton *et al.* showed that a low-carbohydrate diet was sufficient for initial weight loss (6 months), but not effective 12 months into the study [11]. Low-carbohydrate diets often contain high protein. Soenen *et al.* claimed that the high protein component of low-carbohydrate diets is responsible for weight loss [12].

Exercise can prevent the loss of FFM induced by dieting [13]. FFM is the total body mass without the fat. Exercise while on a low-calorie diet increases relative fat loss [13]. Some may shy away from exercise because it increases your appetite, but physical activity increases the satisfaction of a meal [13]. The food you are putting into your body is fuel to complete the activity and the body uses that fuel accurately [13].

Several studies have shown that support is crucial to success in weight loss [14–16]. One study showed face-to-face intervention was better than newsletters or internet-based interventions [14]. Holzapfel *et al.* found no significant correlation between the number of phone calls made to a participant (or the total duration of the calls) and the amount of weight lost within a 12-month period [15]. However, the Drop It At Last (DIAL) pilot study showed that more phone calls with a coach resulted in greater weight loss over a 6-month period [16]. Thus, there are conflicting results on the effect of phone calls.

After initial weight loss, people may have trouble maintaining their weight [17]. Weight re-gain can be caused by a lack of motivation to comply with a diet [17]. Motivation decreases over time and personal motivation is essential for weight loss maintenance [18]. The reward during initial weight loss is witnessing relatively rapid results [17]. Over time and as a person gets closer to their goal, these rewards diminishes as weigh loss slows down [17]. One study showed that one-third of weight loss was regained within a year and the rest within 3-5 years [18]. Another study claimed that on average, overweight individuals lose 5-9% of their original weight in the first 6 months which is followed by weight re-gain [13]. Weight loss counseling can help keep this motivation high and achieve success in long-term weight management.

1.3 PREDICTIVE MODELS

Many weight loss studies examine whether a particular diet or exercise is important while also looking at sex, age, race, family history, and motivational factors [11,17–19]. Some even consider biological factors such as genetic traits [17]. These

studies give insight into influences of weight loss. Fewer studies have used predictive modeling to analyze weight loss.

Two studies developed logistic regression models using early weight loss measurements to determine weight loss success after 12 months [20,21]. Long-term successful weight loss (≥5% loss of body weight by the end of month 12) is associated with age, baseline weight, sex, target caloric intake, and weight loss in previous months [20]. The probability of a participant dropping out of a study can also be modeled with logistic regression [21]. Factors such as lower levels of education and higher levels of obesity contributed to a higher likelihood of dropping out of the program [21]. Sawamoto *et al.* examined predictors of dropout within a particular demographic [22]. A multiple logistic regression was performed on obese Japanese women that took part in a behavioral therapy intervention study [22]. Significant factors contributing to a higher likelihood of dropping out of the study included past mental disorders, greater concern for body image, less organized, the perception of their mothers as less caring, and a higher associated unemployment rate [22]. Logistic regression can generate probabilities of an event at a point in time but methods such as survival analysis can be used to determine the time until that event occurs.

1.4 PROFILE BY SANFORD

Profile by Sanford is a membership based weight loss program that offers one-on-one support [23]. Profile centers its strategy on weight loss coach interactions. Each member has a coach who is available for meetings throughout their time in the program. Coaches help members pick an appropriate meal plan, activity level, educate them, and encourage members to reach their goals.

Members have access to 24 store locations (as of December 2016) where they can meet with a weight loss coach. Profile coaches are trained to guide their members through three phases of the program: Reduce, Adapt, and Sustain [23]. The coach assigns meal plans based on the nutritional needs of the member. These meal plans also follow the reduce, adapt, and sustain philosophy. Meal replacements, shakes, and snacks are available for members to incorporate into their meal plans. Coaches educate individuals on lifestyle changes to help maintain the weight loss they achieve. During one-on-one meetings with their coaches, exercise habits can be discussed as well as eating behaviors that need to be addressed. Coaches usually take the members' waist and hip measurements during their meetings. If a member is not near a Profile store they can utilize the virtual store, which enables telephone communications or video conference with a weight loss coach. Members are encouraged to attend weekly meetings with their coach, either at a store or virtually.

Members also have access to Profile technology (website, smart phone application, and connected scale) which help record daily exercise, food consumption, and weight measurements. Membership includes a body weight scale which links an individual's unique account to the scale. Each time a member weighs themselves, the weight information is sent to Profile servers via Internet. Weight loss progress can be tracked by the individual and their coach to create an effective plan. In this study, we analyze a large set of data gathered for all Profile members using these technologies.

2 METHODOLOGY

2.1 DATA PREPARATION

Data retrieval and data management are vital for statistical analyses. Careful consideration of joining tables and exclusion criteria is crucial. All data mining and preparation were done in R [24]. All figures are generated in R and most are created by the ggplot2 package [25] using the ggplot() function. Some figures are generated by base R graphics functions such as the mosaicplot() and pie() functions.

2.1.1 Data Retrieval

Profile data is stored in a relational database that is hosted by a web server designed to power their website and smart phone application. The web server is accessed through a Virtual Private Network (VPN) connection. This connection allowed for the retrieval of data directly through Structured Query Language (SQL) queries. Such queries are run via the RODBC package [26] directly from R. Each data set was pulled such that the last date of entry was May 31, 2016. The SQL queries were written so that dates were properly formatted and user sensitive information was not pulled. Each desired table was retrieved and saved as a comma-separated value (CSV) file. Table 2-1 lists the eleven tables that were retrieved, along with a brief description.

Table 2-1: Description of individual data sets.

Name	Size (KB)	Description	Rows**	Columns	Section
activity	2,929*	Activity logged such as exercise and the activity intensity and type	136,472	13	2.1.2.1
activity_intensity	1	Description of activity intensity to match the *activity* table	5	4	2.1.2.1
activity_type	1	Description of activity type to match the *activity* table	5	4	2.1.2.1
device_circ	3,124*	Recorded measurements such as hip and waist circumference	417,979	6	2.1.2.2
device_weight	34,437*	Body weight measurements recorded	2,652,106	6	2.1.2.3
food_tag_log	28,837*	Food item logged	3,532,751	7	2.1.2.4
plans	7,357*	Member's meal plans along with start and stop dates	201,876	38	2.1.2.5
store_locations	2	List of each store location; includes store ID	51	3	NA
user_medications	1,072*	Member's disclosed list of medications	87,426	7	2.1.2.6
userinfo	4,041*	Demographic information on each member (excluding sensitive info)	53,451	29	2.1.2.7
userinfo_notes	91,849*	Weight loss coach notes after a meeting with a member	463,784	9	2.1.2.8

*denotes the size of the zipped file
**through May 31, 2016

2.1.2 Cleaning and Formatting of Individual Data Sets

2.1.2.1 Exercise Reported by Members

Any activity or exercise recorded by the user between May 1, 2014 and May 31, 2016 were retrieved from the *activity* data set. This data includes the user ID, date of activity, duration of activity, activity intensity, and activity type. The *activity* data set contains numeric codes for both activity intensity and type. Two tables, *activity_intensity* and *activity_type*, are joined to the *activity* table to obtain the activity type and intensity levels of the recorded exercise. Both activity type and intensity include *Sedentary*, *Light Activity*, *Moderate*, *Very Active*, and *Extra Active* values. After joining these tables, 5805

duplicates were removed from the *activity* table. A monthly summary is generated of the total duration of activity and the number of activities for members with recordings.

2.1.2.2 *Body Size Measurements*

The *device_circ* table contains measurements for areas of the body such as the thigh, hip, waist, chest, bicep, and neck. Each measurement has about 60,000 recordings as shown in Table 2-2. The focus of the *device_circ* data are the hip and waist measurements. Most of these measurements are done by a weight loss coach (99.15%). The other 0.85% of measurements were done with an electronic tape measurement that is no longer utilized in the Profile program. Duplicate recordings based on the measurement and date of the recording were excluded. There was a total of 8811 duplicates removed from the *device_circ* table. Median measurements were collected for each month a member is in the program. A total number of measurements taken for that month is also calculated.

Table 2-2: Number of measurements by body part in the *device_circ* data set.

Body Part	Number of Measurements
thigh	59,463
hip	59,961
waist	60,611
chest	59,978
bicep	59,707
neck	59,612

2.1.2.3 Body Weight Measurements

A total of 448,582 duplicates were removed from the *device_weight* table. Exploratory analysis shows that duplicates from this table were likely due to communication errors between the body weight scale and the database where these

measurements are stored. Time stamps were also removed from the date column; therefore, duplicates could result from a member weighing themselves multiple times within a day with no change in weight. Weight measurements less than 100 pounds and greater than 1000 pounds are excluded, these recordings are considered outlying measurements. Rows that showed a user ID value of 0 were also removed. Since an actual start date is not recorded in any other data sets, the first recorded weight represents the start of the program.

Member's monthly weight measurements were examined to determine if the distribution of the recordings were normally distributed. Some members choose only to record a measurement one to two times a month; those monthly distributions were excluded from this normality test, but not from the data used for further analysis. The result of the Shapiro-Wilks normality test is that 86% of member's monthly weight recording distributions are normally distributed. The results of this analysis justifies using the median monthly weight recording to represent the member's weight for that month. In doing this, any outlying measurements due to other people recording their weight on a member's account will be removed.

A row is generated for each month after a member starts the program to record the date and weight measurement. Each row may or may not contain a weight depending if the member recorded a measurement that month. The data contains the member's ID, the month of the weight recording, weight measurement, months in program (the difference between their start month and the measurement month), and the number of weight measurements they recorded in that month (after removing duplicates).

2.1.2.4 Food Consumption Logged by Members

The *food_tag_log* table contains around 3.5 million observations. Only about 57% of members have used this feature to log food items. The data includes the food item, the date it was consumed, and the number of servings. There are no duplicates in this data. Also included is a meal type ID and a food ID that connect to other data sets to obtain even more information on each food item. These additional data sets include information about which meal the food item was consumed, the color of the food, and nutritional information. Since only a little more than half of the members have utilized this feature, most of this information was not included. The number of food items per month was counted for each member.

2.1.2.5 Meal Plans and Nutrition

The *plans* table contains information about the member's meal plans and dates that the meal plans were utilized. The name of the meal plan, start and end dates, and expected nutritional values for the meal plans are given. The nutritional information includes the calories, protein, carbs, fiber, and activity level. Since the final data is set up on a monthly level by each member and meal plans often change in the middle of the month, a classification process was developed.

Due to some missing values in the field that specify the date a meal plan ended, a new field was created to fill in the missing values. If the meal plan's *date ended* field was missing, but there was a meal plan that started after that meal plan, the date ended would then be the date that the next meal plan started. Otherwise, if there is no following meal plan, the *date ended* is the day that the data was pulled (May 31, 2016); which would denote that the meal plan was the current one being utilized when the data was

retrieved. Any meal plans that have the same start date and end date are eliminated (60584 meal plans were used for zero days) along with 139 duplicates.

Additional rows are added so there is a row for each month that meal plan was used. For example, if a member started a meal plan on June 17, 2014 and ended August 28, 2014, there is a row for June 2014, July 2014, and August 2014 associated with that meal plan for that member. A variable is created to calculate how many days in each month a meal plan was used. From the above example, the meal plan would have been used for 14 days in June, 31 days in July, and 28 days in August. Now, let's say the member started a new meal plan on August 28, 2014 and ended that meal plan on December 2, 2014. They would have used this new meal plan for 4 days in August. Therefore, two rows would be created for the month of August; one would be the first meal plan for 28 days and the second would be the following meal plan for 4 days. In this instance, there are two rows for the month of August and we only want one meal plan to represent a month. The meal plan that was utilized for the most days is chosen to represent that month. In the example, the first meal plan that was used for 28 days in August was selected. If there is a tie between the number of days (each plan was used for 15 days), then the first meal plan that was used in the month is chosen. Finally, each meal plan is grouped into one of eight groups to simplify further analyses. The groups are described in Table 2-3.

Table 2-3: Classification of each meal plan into groups.

Meal Plan Group	Meal Plans
Teen	Teen Recharge, Teen Balance Reduce 1500, Teen Balance Reduce 1200, Teen Balance Adapt 2000, Teen Balance Adapt 2600, Teen Sustain 2000, Teen Sustain 2400
Sustain	Sustain 1500, Sustain 1200, Sustain 1800, 1200 calorie Sustain, 1500 calorie Sustain, Sustain 2000, 1800 calorie Sustain, 2000 calorie Sustain
Jump	Jump Start, Jump concert1
Mom Protocol	Mom Protocol 1700, Mom Protocol 2000, Mom Protocol 2100, Mom Protocol 1800, Mom Protocol 1900, Mom Protocol 2200, Mom Protocol 2500, Mom Protocol 2300, Mom Protocol 2400
Reboot Adapt	Reboot Adapt Step 1 (5'2-5'5), Reboot Adapt Step 1, Reboot Adapt Step 1 (5'6-5'8), Reboot Adapt Week 1, Reboot Adapt Step 1 (5'9-5'11), Reboot Adapt Week 2 - 3, Reboot Adapt Step 2 (5'2-5'5), Reboot Adapt Step 2, Reboot Adapt Step 1 (4'10-5'1), Reboot Adapt Step 2 (5'6-5'8), Reboot Adapt Step 1 (6'0-6'1), Reboot Adapt Step 2 (5'9-5'11), Reboot Adapt Step 1 (6'2-6'3), Reboot Adapt Step 2 (6'0-6'1), Reboot Adapt Step 2 (6'2-6'3), Reboot Adapt Step 2 (4'10-5'1), Reboot Adapt Step 1 (6'4+), Reboot Adapt Step 2 (6'4+)
Reboot Reduce	Reboot Reduce Start, Reboot Reduce (5'2"-5'5"), Reboot Reduce (5'6"-5'8"), Reboot Reduce (5'9"-5'11"), Reboot Reduce (4'10"-5'1"), Reboot Reduce (6'0"-6'1"), Reboot Reduce Optional 3rd Week Beyond (5'2"-5'5"), Reboot Reduce 3rd Week Ongoing, Reboot Reduce Optional 3rd Week Beyond (5'6"-5'8'), Reboot Reduce (6'2"-6'3"), Reboot Reduce Optional 3rd Week Beyond (5'9"-5'11"), Reboot Reduce (6'4"+), Reboot Reduce Optional 3rd Week Beyond (4'10"-5'1"), Reboot Reduce Optional 3rd Week Beyond (6'0"-6'1"), Reboot Reduce Optional 3rd Week Beyond (6'2-6'3"), Reboot Reduce Optional 3rd Week Beyond (6'4"+), Reboot Reduce
Balance	Balanced 1000, Balanced 1200, Balance 1000 (5'3"-5'7"), Balance 1200 (5'3"-5'6"), Balanced 1500, Balance 1200 (5'7"-5'10"), Balance 1000 (4'10"-5'2"), Balance 1500 (5'7"-5'10"), Balanced 1800, Balance 1200 (4'10"-5'2"), Balance 1200 (5'11"-6'1"), Balance 1500 (5'3"-5'6"), Balance 1500 (5'11"-6'1"), Balance 1500 (6'2"+), Balance 1800 (5'7"-5'10"), Balance 1200 (6'2"+), Balance 1800 (6'2"+), Balance 1800 (5'11"-6'1"), Balance 1500 (4'10"-5'2"), Balance 1800 (5'3"-5'6"), Balance 1800 (4'10"-5'2")
Other	Empty Template, Performance 1, NuStart Week 1 – 8, NuStart Week 9 – 12, Research Protocol

2.1.2.6 Medication Indicators for Members

The *user_medications* table contains any medication the member disclosed to their coach. The entries vary from the actual name of the medication to the purpose of the medication and in some cases an abbreviation of either. Keywords were set up to classify medications into 14 different groups. The 14 groups encompass about 76% of the entered medications. The other 24% are medications that did not fall into one of the 14 groups shown in Table 2-4. Each one of the groups in the table is represented as a binary variable in the data where 1 represents the use of that medication and 0 indicates no medication use. In addition to the medication groups, the total number of medications is counted and an indicator of using any medication is created.

Table 2-4: Classification of medication keywords into groups.

Medication Group	Keywords
blood_pressure	blood, pressure, hypertension, lisinopril, bp
antidepressant	depress, anxi, zoloft,
cholesterol	cholesterol
sleep	sleep, insomnia
diabetes	diabet, metformin
thyroid	thyroid, synthroid
acid_reflux	reflux, heartburn, indigest, gerd, acid
vitamin	vitamin, vit, calcium, fish
diuretics	diuretics
alleries	allergies
birthcontrol	birth
asthma	asthma
aspirin	aspirin
bloodthinner	coumadin, warfarin

2.1.2.7 Demographic Data

A comprehensive demographic data set starts with the *userinfo* data. This table contains one row per member and provides information such as zip code, occupation, birthdate, marital status, sex, height, and Profile home store. There are also indicator

columns for members that have been deleted, verified, or those that are Profile coaches. Only records that indicates a verified member is included; deleted observations and Profile coaches are removed.

2.1.2.8 Coach Meeting Descriptions and Notes

Profile coaches utilize the *userinfo_notes* table to enter in notes about their members. Coaches are trained to enter in notes after each meeting. This data is used to determine the number of coach meetings a Profile member attends each month. There are instances where a member is scheduled for a meeting with their coach and does not show up so the coach may enter this information in the *userinfo_notes* table. Since no text mining is performed on the actual message the coach submits, this example would count as a coach meeting when in fact no meeting occurred. From this data, 3574 duplicates were removed. The number of notes entered for each member was counted on a monthly level to represent the number of coach meetings attended in that month.

2.1.3 Merging Individual Data Sets

Combining of the data described in Section 2.1.2 results in an aggregated data set that includes one row per member for each month. This data set is utilized in Chapters 3 and 5 and slightly altered for the use in Chapter 4. The tables are merged one-by-one starting with the *userinfo* table and the cleaned-up *device_weight* table. The *userinfo* table contains those that were indicated as verified members, excluding Profile coaches. Since *userinfo* contains only one row per member, the member's demographic information from this table is repeated for each month. Age was calculated by subtracting the member's birth date from the date in the *device_weight* table and taking the floor of that number. It was discovered that if a member did not specify a birth date,

a default date of January 1, 1970 was inserted into this field. There are 1400 members with this default date. Therefore, any member with a January 1, 1970 birthdate, the value of the age variable was changed to missing (NA) since we do not know their actual age. Cumulative percentage of weight loss was calculated by subtracting the first recorded weight from the current recorded weight and dividing by the first weight. This results in a cumulative percentage of weight loss as a negative number for weight loss and a positive number for weight gain.

Next, information from the *userinfo_notes* table was added. This merge was done on the user's ID and by the month of the aggregated coach meetings. Joining the *userinfo_notes* table adds one column for the number of coach meetings for each month. A zero represents no notes in the *userinfo_notes* table for that month; therefore, no coach meetings.

Body size measurements were added from the *device_circ* table. The actual recording and the number of measurements were joined by the user's ID and the month of the measurement. Additionally, WHR was calculated by dividing the member's waist recording by their hip measurement. An obesity indicator was added based on the WHO standards on the WHR. If a male's WHR is greater than 0.9 or a female's WHR is greater than 0.85, they are considered obese [8].

Summarized activity information is now added to the master table. This includes information taken from the *activity* table such as the monthly total number of activities performed as well as the monthly duration of exercise. If a member does not have any recorded activity for a month, a zero is inserted into the field for the number of activities and the duration of those activities. The number of food items that were logged by a

member each month was added to the existing data. If a member did not record any food items for a month, a zero was input into that field.

Medication use from the *user_medications* table is added with 14 different medication variables along with a medication use indicator and the number of medications used. Since these medications are only disclosed at the beginning of the member's Profile membership and the comprehensive data set contains multiple rows per member, the medication group values are repeated for each month the member is in the program.

Only one meal plan represents a member's monthly plan although multiple meal plans could have been utilized in that month. These representative meal plans are joined with the analysis data. Data added includes the meal plan name, the grouped meal plan, expected calories, protein, carbs, fiber, and activity level, and the plan's start and end dates.

All individual tables are combined into one master table and additional variables are added. An indicator variable is created based on a member being considered *valid* in each month. A valid month for a member occurs when their cumulative percentage weight loss is less than 60 or greater than -60, their monthly percentage of weight loss is less than 15 or higher than -15, the member is between the ages of 18 and 90, and their meal plan is not in the Mom Protocol or Teen groups. If any of these criteria are violated, the observation is considered invalid.

BMI is calculated each month as well as a starting BMI variable. BMI is calculated in Equation 2-1,

$$\text{BMI} = \left(\frac{weight\ (lbs)}{height(in)^2}\right) * 703. \qquad 2\text{-}1$$

Members can be classified into groups based on their BMI [6]. A variable is created to classify members into categories found in Table 2-5.

Table 2-5: BMI classification groups.

BMI Group	BMI Range
Underweight	< 18.5
Normal	18.5 - 24.9
Overweight	25 - 29.9
Obese I	30 - 34.9
Obese II	35 - 39.9
Obese III	≥ 40

A cumulative count of coach meetings is calculated for each member. Multiple variables are created to represent behavior from the member's previous month. These variables include the member's previous month's weight, cumulative percentage of weight loss, meal plan, number of weight recordings, number of coach meetings, cumulative number of coach meetings, BMI, BMI group, number of activities logged, number of food items recorded, waist measurement, hip measurement, number of waist measurements, number of hip measurements, WHR, WHR obesity indicator, and number of pounds lost.

2.2 STATISTICAL TECHNIQUES

Predictive models are utilized in many industries to predict a future outcome based on given characteristics. Models can be used in health care to predict emergency room volumes; financial companies can forecast customer default rates; call centers can forecast call volumes; insurance agencies can assign risk levels to policyholders;

customer centric companies can forecast customer attrition. These are only a few examples of where predictive modeling can help organizations increase efficiency and profitability.

2.2.1 Basis Spline Functions

Spline functions are often used to fit a curve without a parametric form. Spline functions generate spline curves which are piecewise polynomial curves that fit together [27]. A spline function, by definition, is a linear combination of n B-splines, $B_{j,d}(x)$, of order d with knot sequence $\boldsymbol{t} = (t_j)$ [27,28]. The spline function, which is constructed from n control points $(c_j)_{j=1}^{n}$, can be written as

$$f = \sum_{j=1}^{n} c_j B_{j,d}\,. \tag{2-2}$$

The n control points $(c_j)_{j=1}^{n}$ are also considered the B-spline coefficients of the function f [27]. The number of control points, or consequently the number of B-splines, is equal to the degree of the spline function plus the number of internal knots [29]. The knot sequence, $\boldsymbol{t}$, is defined as a non-decreasing sequence of real numbers that is $n + d + 1$ in length [27]. The jth B-spline, for all real numbers x, is defined as

$$B_{j,d}(x) = \frac{x - t_j}{t_{j+d} - t_j} B_{j,d-1}(x) + \frac{t_{j+1+d} - x}{t_{j+1+d} - t_{j+1}} B_{j+1,d-1}(x), \tag{2-3}$$

where

$$B_{j,0} = \begin{cases} 1, & \text{if } t_j \le x < t_{j+1} \\ 0, & \text{otherwise.} \end{cases} \tag{2-4}$$

Examples in this section will refer to time as the independent variable in which the spline function is defined on. For simplicity, discussion of splines will be limited to examples where a knot is placed at time 6 with boundary points at time 0 and 12, the knot sequence then depends on the degree of the spline function.

First, we will start with a simple spline function of degree 0 where our knot sequence can be defined as $\boldsymbol{t} = (t_1, t_2, t_3) = (0, 6, 12)$. Since $d = 0$ we can utilize Equation 2-4 with only one basis function,

$$B_{1,0}(x) = \begin{cases} 1, & \text{if } t_1 \leq x < t_2 \\ 0, & \text{otherwise.} \end{cases}$$

The basis function is graphically represented in Figure 2-1.

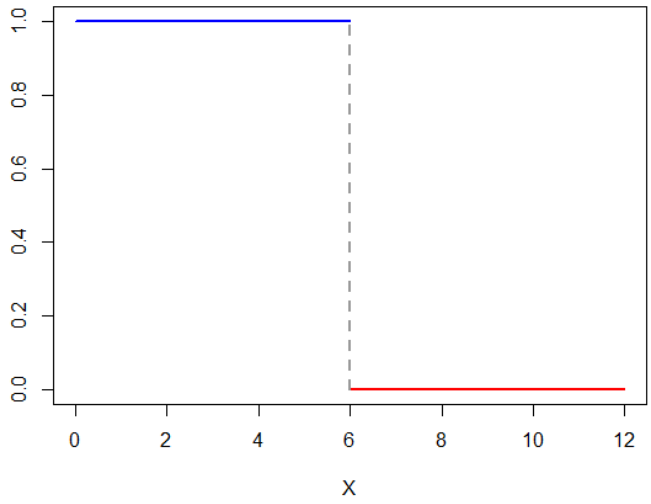

Figure 2-1: Representation of a spline function of degree 0 and one internal knot at 6.

With degree 0, this spline function consists of one basis function and is represented as:

$$f = \begin{cases} 1, & if\ 0 \leq x < 6 \\ 0, & \text{otherwise.} \end{cases}$$

Adding complexity with a first-degree spline, two basis functions are formed where our knot sequence can be defined as $\boldsymbol{t} = (t_1, t_2, t_3, t_4) = (0, 6, 12, 12)$ based on Equation 2-3,

$$B_{1,1}(x) = \frac{x - t_1}{t_2 - t_1} B_{1,0}(x) + \frac{t_3 - x}{t_3 - t_2} B_{2,0}(x), \qquad 2\text{-}5$$

$$B_{2,1}(x) = \frac{x - t_2}{t_3 - t_2} B_{2,0}(x) + \frac{t_4 - x}{t_4 - t_3} B_{3,0}(x). \qquad 2\text{-}6$$

By utilizing Equation 2-4, $B_{1,0} = 1$ if $0 \leq x < 6$, $B_{2,0} = 1$ if $6 \leq x < 12$, and $B_{3,0} = 0$ everywhere. Therefore, Equations 2-5 and 2-6 become

$$B_{1,1}(x) = \begin{cases} \frac{x}{6} & \text{if } 0 \leq x < 6 \\ \frac{12 - x}{6} & \text{if } 6 \leq x < 12, \end{cases} \qquad 2\text{-}7$$

$$B_{2,1}(x) = \begin{cases} 0 & \text{if } 0 \leq x < 6 \\ \frac{x - 6}{6} & \text{if } 6 \leq x < 12. \end{cases} \qquad 2\text{-}8$$

These two functions are represented graphically in Figure 2-2.

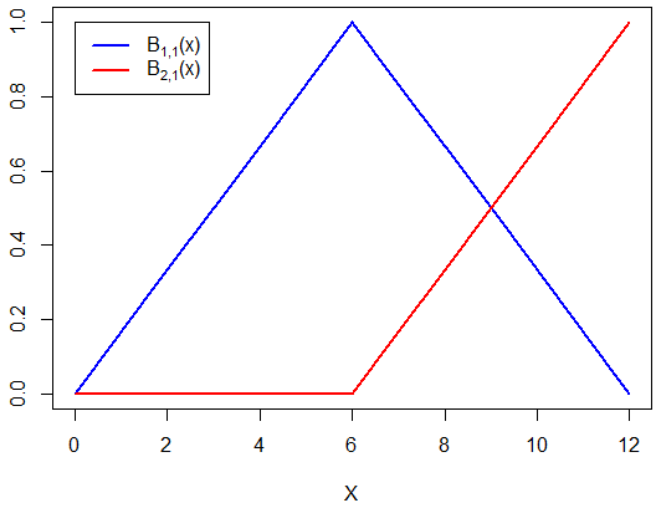

Figure 2-2: Representation of a first-degree spline function with an internal knot at 6.

Since the spline function is a linear combination of B-splines, we can describe this spline function of degree 1 as shown in Equation 2-2,

$$f = \begin{cases} c_1 \dfrac{x}{6} & \text{if } 0 \le x < 6 \\ c_1 \dfrac{12 - x}{6} + c_2 \dfrac{x - 6}{6} & \text{if } 6 \le x < 12, \end{cases} \qquad 2\text{-}9$$

where c_1 and c_2 are the B-spline coefficients.

Constructing a quadratic spline function is the same process as above. Graphically, the three basis functions are shown in Figure 2-3.

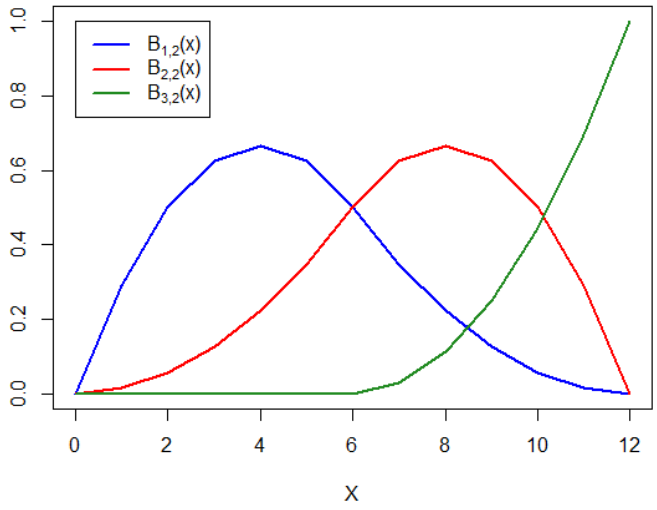

Figure 2-3: Representation of a second-degree spline function with an internal knot at 6.

Thus, a second-degree spline function with boundary knots at 0 and 12 and an internal knot at 6 is written as,

$$f = \begin{cases} c_1 \dfrac{(12 - x)^2}{72} + c_2 \dfrac{x^2}{72} & \text{if } 0 \le x < 6 \\ c_1 \dfrac{-x^2 + 8x}{24} + c_2 \dfrac{-x^2 + 16x - 48}{24} + c_3 \dfrac{(x - 6)^2}{36} & \text{if } 6 \le x < 12. \end{cases} \qquad 2\text{-}10$$

Additional internal knots can be added to the function which increases the number of basis functions. The same process would be followed to generate the spline function with additional knots.

2.2.2 Linear Mixed Effects Models

If individuals have multiple measurements of a covariate over time, a linear mixed effects model is often used [30]. A mixed effects model contains both fixed and random effects [31]. Fixed effects are generally referred to as the average population effect and random effects are subject-specific [30].

The modeling process is based on the idea that each individual has their own subject-specific mean response profile [30]. A basic representation of a mixed effects model for response, y_{ij}, is

$$y_{ij} = (\beta_0 + b_{i0}) + (\beta_1 + b_{i1})t_{ij} + \varepsilon_{ij}, \qquad 2\text{-}11$$

where β_0 and β_1 are considered fixed effects and represent the individual's average intercept and slope, respectively [30]. Random effects, b_{i0} and b_{i1}, represent the deviation from the average intercept and slope for individual i [30]. We also let t_{ij} represent time for individual i, $i = 1, \dots, n$; $j = 1, \dots, n_i$ for n-subjects and the error terms ε_{ij} are assumed to come from a normal distribution with mean zero and variance σ^2 [30]. A generalization of the linear mixed effects model has the form:

$$\begin{cases} y_i = X_i\beta + Z_i b_i + \varepsilon_i, \\ b_i \sim \mathcal{N}(0, D), \\ \varepsilon_i \sim \mathcal{N}\left(0, \sigma^2 I_{n_i}\right), \end{cases} \qquad 2\text{-}12$$

where X_i and Z_i are known design matrices of fixed and random variables, β is a vector of fixed parameters, b_i is a vector of random effects, and I_{n_i} denotes the n_i-dimensional identity matrix [30]. The random effects, b_i, are assumed to be independent of the error terms ε_i with mean zero and variance-covariance matrix D [30]. Interpretation of the fixed effects, β, is the same as a simple linear regression; β denotes the change in the average y_i with one unit increase in the covariate associated with β when all other covariates are held constant [30,31]. The random effects, b_i, can be interpreted as the deviation of the i^{th} subject from the average, β [30,31].

Simple linear regression applies the same intercept and slope to each subject. A mixed effects model allows varying intercept and slopes for each subject [30]. Additionally, mixed effects models allow for missing response data and does not require the same number of observations per subject [30].

2.2.3 Survival Analysis

Survival analysis is utilized in several different types of analyses. For predictive modeling, where we are interested in determining the probability of an event occurring after a particular time, a semiparametric or parametric model is needed. A survival function is used to describe this probability that the event occurs after time t or alternatively, the probability of surviving to time t,

$$S(t) = pr(T > t), 0 < t < \infty, \qquad 2\text{-}13$$

where T represents the random variable of failure times [32]. Survival functions can be defined in terms of the hazard function. The hazard function describes the instantaneous

failure rate or the risk of an event within $[t, t+dt]$ provided that the subject survived to time t [30]. This can also be referred to as the risk function and defined as

$$\mathrm{h(t)} = \lim_{dt \to 0} \frac{\mathrm{pr}(t \leq T < t+dt \mid T \geq t)}{dt}, \mathrm{t} > 0. \qquad 2\text{-}14$$

The complement of the survival function,

$$F(t) = pr(T \leq t), 0 < t < \infty, \qquad 2\text{-}15$$

is commonly known as the cumulative distribution function (CDF) or cumulative risk function in survival analysis [32]. Therefore, the probability density function (PDF) is defined as

$$f(t) = \frac{d}{df}F(t) = -\frac{d}{df}S(t). \qquad 2\text{-}16$$

We can now use the PDF and survival function equations to define the hazard function as,

$$h(t) = \frac{f(t)}{S(t)}. \qquad 2\text{-}17$$

We can also define a cumulative hazard function which is the area under the hazard function up to time t as,

$$H(t) = \int_0^t h(u)du\,, \qquad 2\text{-}18$$

and finally give the survival function in terms of the hazard function,

$$S(t) = exp\left(-\int_0^t h(u)du\right) = exp\big(-H(t)\big)\,. \qquad 2\text{-}19$$

2.2.3.1 Nonparametric Models

The Kaplan-Meier estimate is common when discussing nonparametric survival methods. The Kaplan-Meier estimate for the survival function is

$$\hat{S}(t) = \prod_{t_i \leq t} \left(1 - \frac{d_i}{n_i}\right), \qquad 2\text{-}20$$

where n_i represents the number of subjects at risk at time t_i and d_i represents the number of events at time t_i [32]. This estimate contains no assumed parametric distribution. Nonparametric survival methods are particularly useful when we want to compare survival curves of two groups, such as an experimental group and control group [32]. Nonparametric methods will be examined as an exploratory analysis, but since this method is not able to generate survival probabilities, other methods are utilized more extensively.

2.2.3.2 Parametric Models

Parametric survival models are based on a distribution for the hazard function, $h(t)$ [32]. A simple survival distribution is the exponential distribution which has a constant hazard, $h(t) = \lambda$ [32]. We can derive the cumulative hazard function by referencing Equation 2-18,

$$H(t) = \int_0^t h(u)du = \int_0^t \lambda du = \lambda t \,|_0^t = \lambda t.$$

Consequently, we have a survival function of $S(t) = e^{-\lambda t}$ and PDF of $f(x) = \lambda e^{-\lambda t}$. Several other distributions can be utilized for a parametric survival model depending on the distribution that best fits the data.

Unlike nonparametric survival models, parametric models do generate a survival probability based on covariates. Parametric models lack the flexibility to capture the shape of the hazard function and patient-specific survival predictions are highly dependent on a correct baseline hazard function [29].

2.2.3.3 Semiparametric Models

A proportional hazards model stems from the previous idea of wanting to examine the difference between two survival distributions. This difference can be defined using the parameter, ψ, in what is known as the Lehmann alternative, $S_1(t) = S_0(t)^{\psi}$ [32]. Utilizing the relationship between the survival function and the hazard function we know that $h_1(t) = \psi h_0(t)$ [32]. This association is known as the proportional hazards assumption [32]. We can also allow the inclusion of covariates in vector z by letting $\psi = e^{z\beta}$ [32]. There are no assumptions made about the distribution of event times with a proportional hazards model [30]. The partial log-likelihood function does not require a baseline hazard to be specified [30]. Instead, the model assumes that covariates act multiplicatively on the hazard rate [30].

Cox proportional hazards model is a semiparametric model that extends the proportional hazards model by using the partial likelihood function [33]. The partial likelihood allows for a baseline survival distribution to be defined by covariates instead of a specific parametric survival distribution [32]. A basic representation of the Cox proportional hazards model is,

$$h_i(t|w_i) = h_0(t)e^{\gamma^T w_i}, \qquad 2\text{-}21$$

where $h_0(t)$ is an unspecified baseline hazard function, γ is a vector of regression coefficients, and $w_i^T = (w_{i1}, w_{i2}, \dots, w_{ip})$ is a vector of covariates [30]. Taking the log of Equation 2-21,

$$log\big(\mathrm{h_i(t|w_i)}\big) = \log\big(\mathrm{h_0(t)}\big) + \gamma_1 \mathrm{w_{i1}} + \gamma_2 \mathrm{w_{i2}} + \cdots + \gamma_p \mathrm{w_{ip}},$$

γ_j is described as the change in the log hazard at any time t with one unit increase of w_j with all other predictors held constant [30]. Similarly, e^{γ_j} is the ratio of hazards for a unit change in the corresponding covariate [30]. Comparing hazards of subject i and subject j, we would have the following ratio,

$$\frac{h_i(t|w_i)}{h_j\big(t|w_j\big)} = \frac{h_0(t)e^{\gamma^T w_i}}{h_0(t)e^{\gamma^T w_j}} = \frac{e^{\gamma^T w_i}}{e^{\gamma^T w_j}} = e^{\gamma^T (w_i - w_j)}, \qquad 2\text{-}22$$

where the baseline hazard function no longer exists in the equation [30]. Equation 2-22 represents the hazard ratio for subject i compared to subject j [30]. The Cox proportional hazards model is considered semiparametric since the baseline hazard does not assume a parametric form but the covariates are in the model linearly [33]. Due to the ability to generate predicted survival probabilities, incorporate covariates into this prediction, and its flexibility; a semiparametric survival model is used in this research.

2.2.3.4 Time-Dependent Covariates and Extended Cox Models

The partial likelihood method applied to survival data allows for the inclusion of covariates to model survival times [32]. An assumption of this inclusion is that covariates are measured at baseline and do not change [32]. Covariates such as sex, starting weight, age at the beginning of the study, or occupation fit this assumption. Many relevant covariates do not remain constant throughout the study though. Time-

dependent covariates such as an individual's weight or blood pressure could be important factors in the study as well as the season or month of the year. There are two categories of covariates that change over time. An individual's weight or blood pressure at any time t is unknown and referred to as an endogenous time-dependent covariate [30]. The month of the year at any time t is known and referred to as an exogenous time-dependent covariate [30].

Exogenous variables are usually measured without error, predictable and known at any time t before time t occurs [30]. An event at time s, where $t > s$, does not affect the value of the exogenous variable at time t [30]. Spring will always start in March and end in June (in the Northern Hemisphere) even if the event of interest occurs within that time.

Endogenous variables are measured with some error, not predictable and typically if an event occurs they can no longer be measured [30]. For example, if the event is death and the endogenous variable is blood pressure measurements, once death occurs the patient's blood pressure can no longer be recorded. These measurements are only known at measurement times and their complete path to time t is not fully observed [30]. An individual's blood pressure can change from one hour to the next whereas measurements for a study might only be recorded weekly.

A Cox proportional hazards model assumes that covariates are constant between follow-up times [30]. This is true for variables such as sex, a specific treatment, or any baseline measurements. The problem arises when we want to include time-dependent covariates. Adjustments are required to obtain unbiased estimates in order to include time-dependent variables in a Cox proportional hazards model [32]. To use the Cox

model, we would need to know the values of the time-dependent covariate as a time continuous process without measurement error to maximize the partial likelihood in order to estimate the parameters [34]. An adjustment can be made which modifies the partial likelihood function and yields the extended Cox model where exogenous time-dependent covariates can be utilized [32].

An extended Cox model can be utilized if encountering exogenous time-dependent covariates [30]. This method assumes that covariates are predictable and measured without error [30]. As stated above, endogenous covariates are unknown for future times and measurements such as body weight and blood pressure carry a certain amount of measurement error. The extended Cox model is not appropriate to use with endogenous time-dependent covariates.

The partial likelihood method to estimate a parameter of a time-dependent covariate requires a measurement for every uncensored event time [35]. Most often, an individual's measurements such as blood pressure are measured irregularly over time and therefore the partial likelihood method is not applicable [35]. Imputation is sometimes utilized in which the last observation is carried forward to account for a missing measurement at an observed time event [35]. This imputation method can introduce bias into the parameter estimations. Additionally, these type of measurements often come with measurement error and may not truly reflect observed values [35].

Another alternative to the Cox model and the extended Cox model is a two-stage modeling approach [30,34]. First, the longitudinal process is modeled using a least-squares method which is then used to impute these values into the partial likelihood for the Cox model and the partial likelihood is then maximized [30,34]. This method reduces

the parameter estimate bias in the Cox model but is still not an unbiased approached [30,34]. The two-stage approach does not utilize any survival information when modeling the longitudinal process whereas the joint likelihood method uses the survival and longitudinal data simultaneously [30,34].

2.2.4 Joint Modeling

Joint modeling is an enhancement of survival analysis which associates the prediction of a longitudinal measure with a time to an event [34]. Most commonly, in biostatistics, a biomarker that is repeatedly measured over time may be predictive of an event such as death or onset of a disease. As discussed in the previous section, the Cox proportional hazards model does not allow for the inclusion of such endogenous covariates [30,36], which is where joint modeling plays a significant role in this research. Joint modeling reduces bias of other proposed methods such as the extended Cox model and the two-stage model and improve predicted survival probabilities [37].

Joint modeling allows for the inclusion of time-dependent covariates to model the time to an event, but it also can be utilized to associate the relationship between the covariate and risk of the event [30]. Additionally, joint modeling has been used to determine a surrogacy to an event such as cancer biomarkers so that a biomarker can be an indicator of cancer progression or regression [35].

2.2.4.1 Longitudinal Submodel

The longitudinal submodel is a linear mixed effects model. As mentioned in Section 2.2.2, $y_i(t)$ denotes the observed value of the longitudinal outcome. The predicted outcome for the linear mixed effects model is denoted by $m_i(t)$. The observed

longitudinal value, $y_i(t)$, is the true outcome, $m_i(t)$, plus a random error, $\varepsilon_i(t)$ [30]. We can express this model as,

$$\begin{cases} y_i(t) = m_i(t) + \varepsilon_i(t), \\ m_i(t) = x_i^{\mathrm{T}}(t)\beta + z_i^{\mathrm{T}}(t)b_i, \\ b_i \sim \mathcal{N}(0, D), \qquad \varepsilon_i(t) \sim (0, \sigma^2), \end{cases} \tag{2-23}$$

where $x_i(t)$ are fixed effects with parameters β and $z_i(t)$ are random effects with parameters b_i [30].

2.2.4.2 Joint Model

As discussed in Section 2.2.3.4, utilizing time-dependent endogenous covariates in the Cox model, the extended Cox model, or the two-stage model either violate critical assumptions or introduce parameter bias. The method of joint modeling alleviates this bias and improves survival predictions [34]. As opposed to the two-stage modeling approach, joint modeling uses the likelihood method based on maximizing the log-likelihood of the joint distribution of both the survival and longitudinal data [38]. Thus, the survival and longitudinal data are used simultaneously. This approach assumes that the random effects account for the correlation between the longitudinal repeated measures as well as the association between the longitudinal outcome and the survival events; the random effects are shared between the two processes [38].

Joint models extend the Cox proportional hazards model and have a similar functional form as the extended Cox model. The notation for the Cox model is shown in Equation 2-21. To incorporate the longitudinal outcome into the model, we include the current value, $m_i(t)$, into our hazard function,

$$h_i(t|\mathcal{M}_i(t), w_i) = h_0(t)\, exp\{\gamma^T w_i + \alpha m_i(t)\}, t > 0. \tag{2-24}$$

It is important to note that $\mathcal{M}_i(t)$ refers to the entire longitudinal process up to time point t; $\mathcal{M}_i(t) = \{m_i(s), 0 \leq s < t\}$ [30]. We interpret α as the relationship between the risk of an event and the current longitudinal outcome [30].

As previously discussed, the survival function can be expressed in terms of the hazard function,

$$S(t) = exp\left(-\int_0^t h(u)du\right). \quad 2\text{-}25$$

Therefore, we can write our joint model survival function as,

$$S(t|\mathcal{M}_i(t), w_i) = exp\left(-\int_0^t h_0(s)\, exp\{\gamma^T w_i + \alpha m_i(s)\}\, ds\right). \quad 2\text{-}26$$

This survival function implies that our survival probability depends on $\mathcal{M}_i(t)$, the entire longitudinal history up to time t [30]. Whereas the hazard function in Equation 2-24 only depends on the current longitudinal outcome at time t, $m_i(t)$ [30]. Extensions of the model have been developed to further integrate this outcome.

The joint hazard model assumes that the risk of dropping out of the program in month t depends on the predicted longitudinal outcome in that same month. The first extension involves incorporating the slope of the longitudinal outcome,

$$h_i(t) = h_0(t)\, exp\{\gamma^T w_i + \alpha_1 m_i(t) + \alpha_2 m_i'(t)\}, \quad 2\text{-}27$$

where $m_i'(t) = \frac{d}{dt} m_i(t)$. Interpretation of α_1 is the same as α in Equation 2-24. Since $m_i'(t)$ represents the slope of the longitudinal outcome over time, α_2 is the relationship between the slope and the risk of an event at time t when the current outcome, m_i, is held

constant. For example, if two members have the same outcome at time t but one has a positive slope and the other has a negative slope of their longitudinal trajectory, we may expect different outcomes in event risk. Figure 2-4 illustrates this difference with arbitrary member behavior.

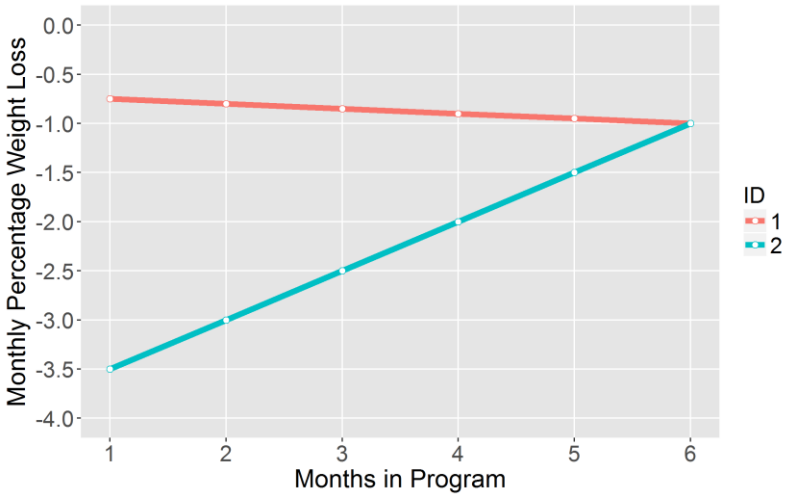

Figure 2-4: Example of differences in longitudinal trajectory over time.

The second extension considers the cumulative value of the longitudinal outcome which can be represented in the following hazard function,

$$h_i(t) = h_0(t)\, exp\left\{\gamma^T w_i + \alpha_1 m_i(t) + \alpha_2 \int_0^t m_i(s)ds\right\}. \qquad 2\text{-}28$$

Instead of the model depending only on the current value of the longitudinal outcome, it depends on the cumulative value to time t calculated by the integral of $m_i(t)$. In this extension, α_2 represents the association between the cumulative value of the longitudinal outcome and the risk of an event.

Unlike the Cox model in which a baseline hazard function is not specified, the joint models described above need a specified baseline hazard function, $h_0(\cdot)$ [30]. It is shown by Rizopoulos that an unspecified baseline hazard function will underestimate the

standard errors of the parameter estimates [30]. Therefore, a parametric model for $h_0(t)$ is ideal. Parametric distributions such as exponential or Weibull are less flexible and utilizing a spline function or piecewise-constant function can be more flexible. Patient-specific survival predictions rely heavily on a correct baseline hazard function, therefore a flexible but accurate depiction is crucial [29]. For example, the piecewise-constant baseline hazard function looks like this:

$$h_0(t) = \sum_{q=1}^{Q} \xi_q I\left(\upsilon_{q-1} < t \leq \upsilon_q\right), \qquad 2\text{-}29$$

where $0 = \upsilon_0 < \upsilon_1 < \cdots < \upsilon_Q$ are points in time, with υ_Q being larger than the largest observed time, and ξ_q is the hazard within $(\upsilon_{q-1}, \upsilon_q]$ [30]. The piecewise-constant function contains $(Q-1)$ internal knots and as the number of knots increase, the more flexible the baseline hazard [30].

Joint models with flexible baseline hazard functions estimate parameters by utilizing a joint distribution to alleviate bias. This method has improved on previous methods that wrongly employed endogenous variables and methods that introduce bias into the parameter estimates [37]. Joint modeling improves this process and while it may be computational extensive, has been used to improve prediction and reduce bias [37].

3 EXPLORATORY DATA ANALYSIS

3.1 ANALYSIS OF INDIVIDUAL DATA SETS

Demographic data provide basic member information. Available data such as body weight, hip, and waist measurements, activity levels, food intake, meal planning, medication use, and weight loss coach interactions are all combined with this basic information to give insight into Profile member's weight loss journey. Each separate table is described below before combining the data into one master data set to be analyzed at a monthly level. Analysis of demographic information such as age, sex, and marital status is performed in combination with other data. Demographic data is combined with data tables described below in Section 3.1.1 through Section 3.1.7 and discussed in Section 3.2.

3.1.1 Body Weight

When joining Profile, each member receives a scale to measure their body weight. The scale is linked to the member's account via WiFi connection. This functionality allows members to easily track their body weight. Measurements are stored in a table that records the user's weight measurement (in pounds) and the time of the measurement. Members can record their weight at any time during the day. Some members have measured their weight up to 12 times in one day. Body weight can also be recorded manually during a member's meeting with a weight loss coach. An indication of manual or scale recording is documented in the data.

Figure 3-1 shows that the total number of weight recordings each month increases from month to month. This increase is due to an increase in the number of Profile

members recording their weight. Figure 3-2 shows that the average number of monthly weight recordings for members is consistently between 7 to 8 measurements each month.

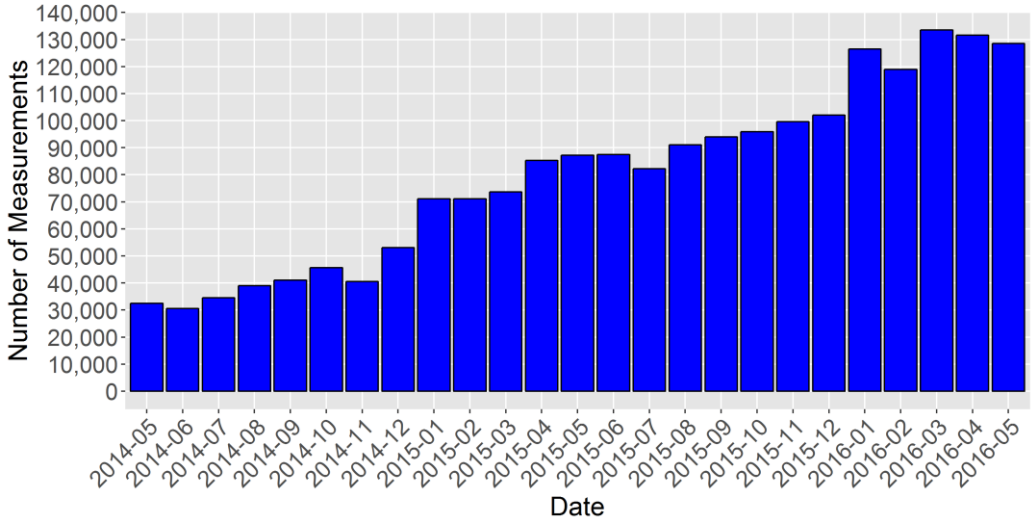

Figure 3-1: Total number of weight measurements recorded by the year and month.

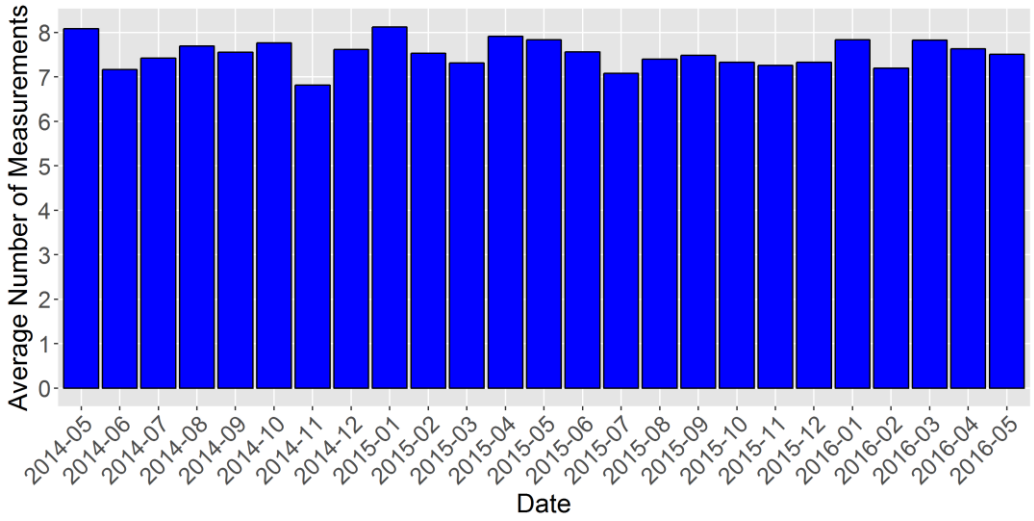

Figure 3-2: Average number of weight measurements per members by the year and month.

The distribution of weight recordings by the hour in which the measurements were taken is shown in Figure 3-3. 22% of measurements are recorded in the 6 o'clock hour. In general, weight measurements are being recorded in the morning with 68% after 5 AM and before 9 AM. All time stamps are converted to central time zone.

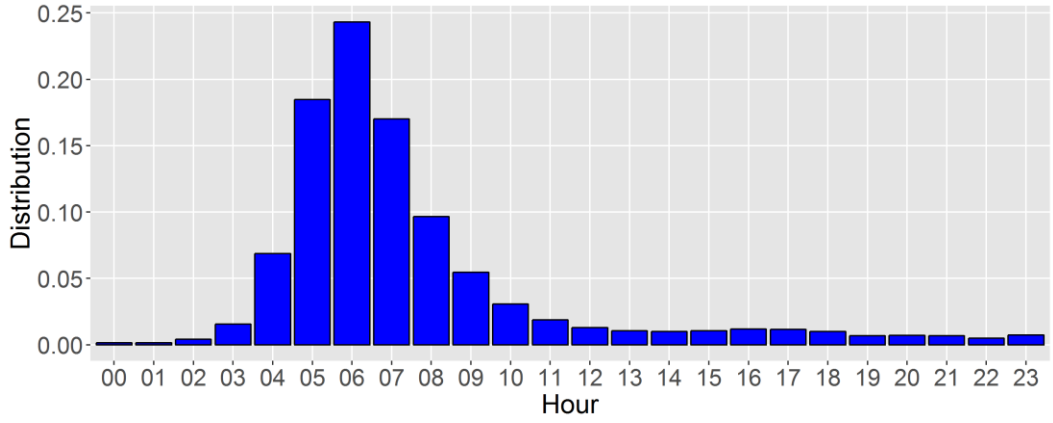

Figure 3-3: Distribution of weight recordings by the hour in which weight was measured.

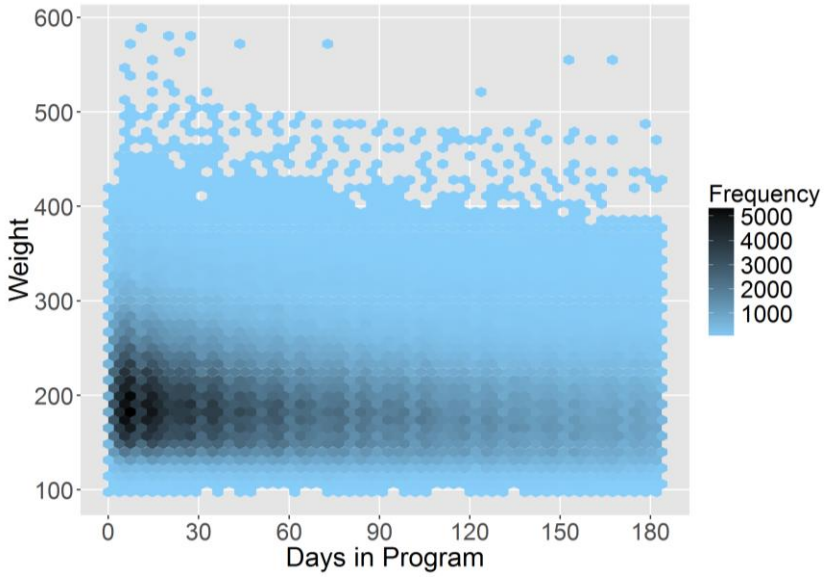

Figure 3-4: Weight measurements within the first six months in the program. Darker areas represent more measurements.

Figure 3-4 illustrates weight measurements within the first six months in the program. This graph contains 1,383,306 weight measurements. The darker areas denote a higher concentration of measurements whereas the light blue areas denote less measurements. The graph indicates that most measurements are between 150 and 250 pounds and as time goes on fewer measurements are being recorded.

3.1.2 Coach Meeting

Profile members are given the opportunity to attend one-on-one meetings with a weight loss coach. These meetings occur in-person at a Profile store or by a virtual meeting over the phone. The data does not distinguish between in-person or phone interaction. During these meetings members set up a plan to achieve their goals and discuss their progress. Profile encourages members to meet with their coach once a week early in the program and continue with meetings throughout their membership. Figure 3-5 shows that members are attending an average of two meetings a month. This graph does not take into consideration members that are not attending coach meetings. It will be shown in a later section that several members choose to attend zero meetings a month.

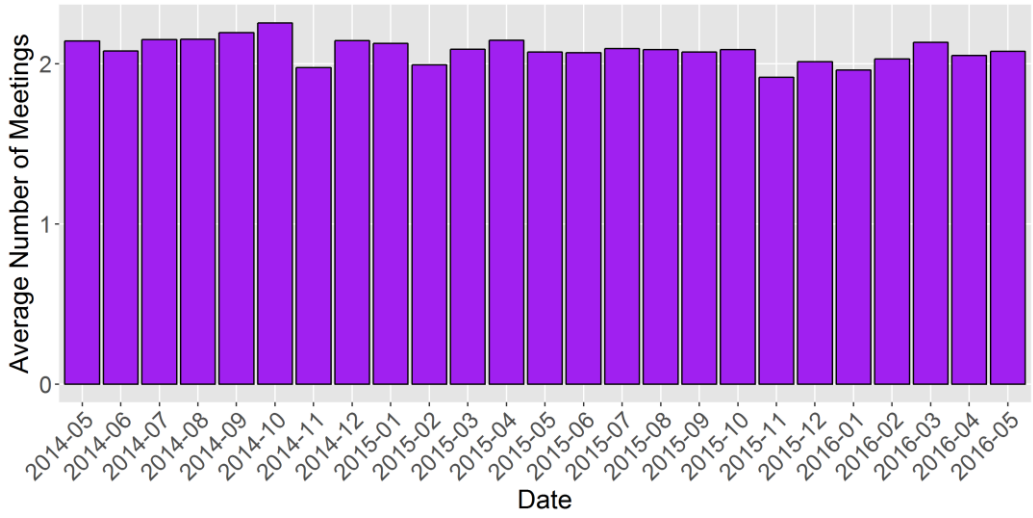

Figure 3-5: Average number of monthly coach meetings per member by year and month (by members attending meetings).

3.1.3 Hip and Waist Measurements

Measurements of hip and waist are recorded less frequently than body weight measurements. An electronic device to measure hip and waist circumference was utilized early in the program development but has been discontinued. Some of the recorded

measurements were done with the electronic device and some were manually recorded during a meeting with a coach. Waist and hip measurements over 200 inches were considered outliers. The distribution of hip measurements is shown in Figure 3-6 while the distribution of waist measurements is shown in Figure 3-7. Most hip measurements (93%) are between 35 and 55 inches and most waist measurements (88%) are between 30 and 50 inches. The average WHR for these measurements is 0.88 which would be considered obese for females but not for males.

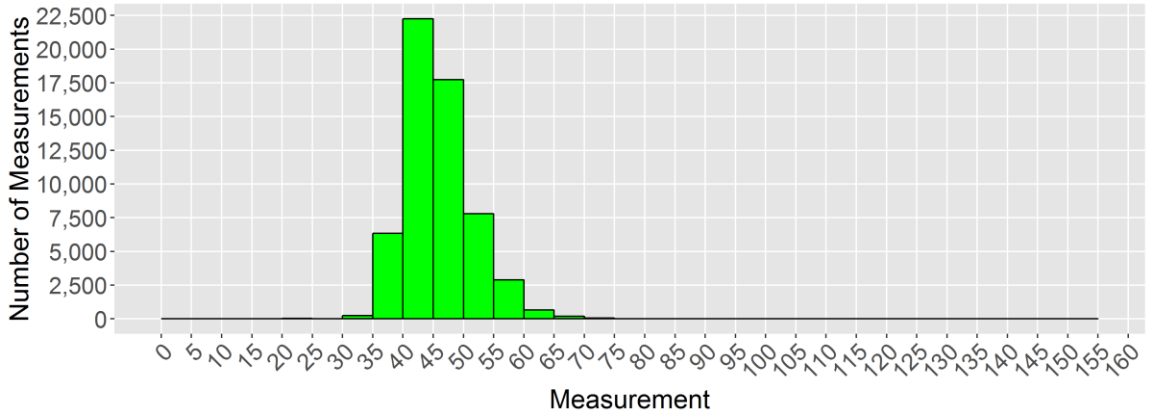

Figure 3-6: Distribution of hip measurements in inches.

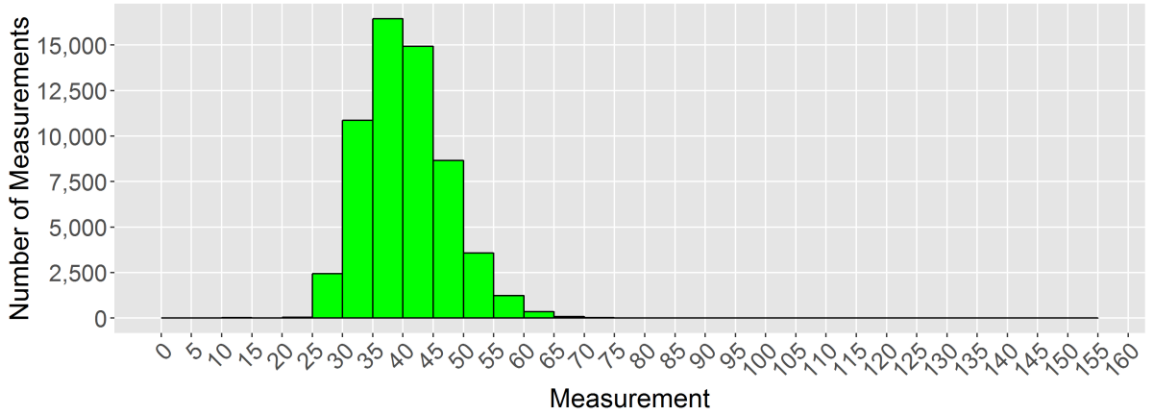

Figure 3-7: Distribution of waist measurements in inches.

3.1.4 Physical Activity and Exercise

Only 20% of members have registered any activity. Activities are classified by their intensity and type based on the member's perception of the activity. Table 3-1 shows the distribution of these activity classifications. The largest percent (40%) of activity is a sedentary activity performed with moderate intensity.

Table 3-1: Distribution of activity classification by activity type and activity intensity.

Activity Type	Activity Intensity				
	Sedentary	Light Activity	Moderate	Very Active	Extra Active
Sedentary	6%	12%	40%	20%	3%
Light Activity	1%	3%	2%	0%	0%
Moderate	1%	2%	6%	4%	0%
Very Active	0%	0%	0%	0%	0%
Extra Active	0%	0%	0%	0%	0%

The duration of activity is recorded and summarized in Figure 3-8. Each recorded activity is 60 minutes or less and out of all recorded activity 22% are 30 minutes in length, and 57% are 30 minutes or less. Most members are logging less than 60 minutes of total monthly activity. Figure 3-9 is truncated at 1000 minutes of exercise although 4% of members have recorded more than 1000 minutes of exercise in a month.

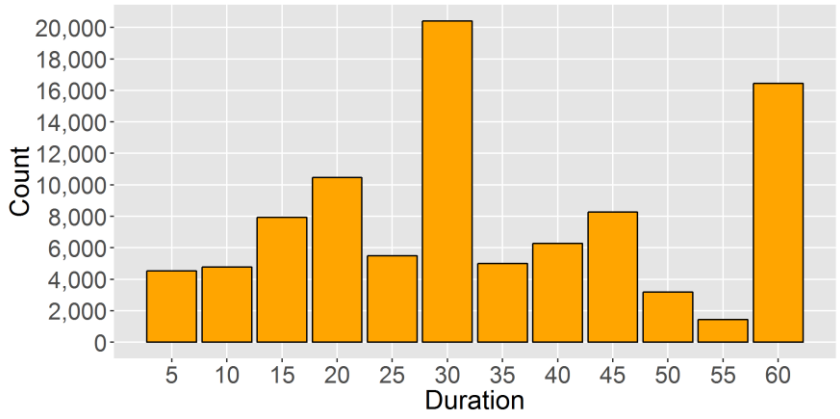

Figure 3-8: Distribution of each logged activity length.

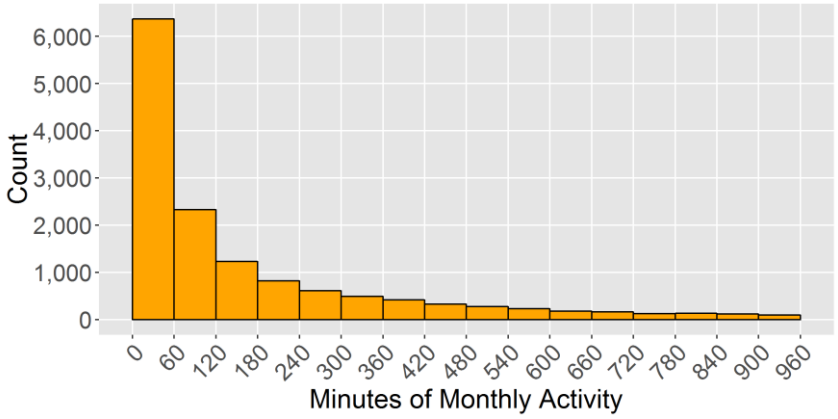

Figure 3-9: Distribution of total duration of logged monthly activity.

3.1.5 Food Items Logged

Food items can be logged by a Profile member. Similar to activity recordings, this feature is not utilized by all members. 57% of participating members have documented at least one food item. If a member is recording their food intake, several items are logged for one meal which makes this data table large. There are almost 2,751,684 food items recorded since May 1, 2014. Figure 3-10 shows the distribution of the daily food items logged where 91% of recording are 12 or fewer items in a day. Each January there is a slight increase in the number of logged items, as shown in Figure 3-11.

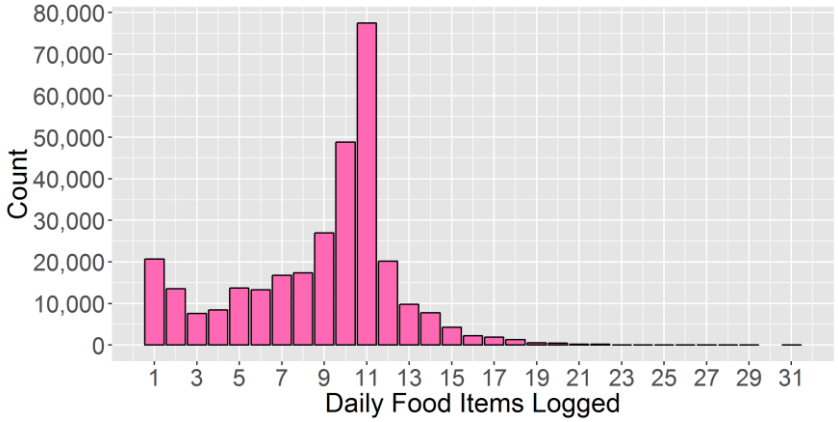

Figure 3-10: Distribution of the number of food items logged in a day.

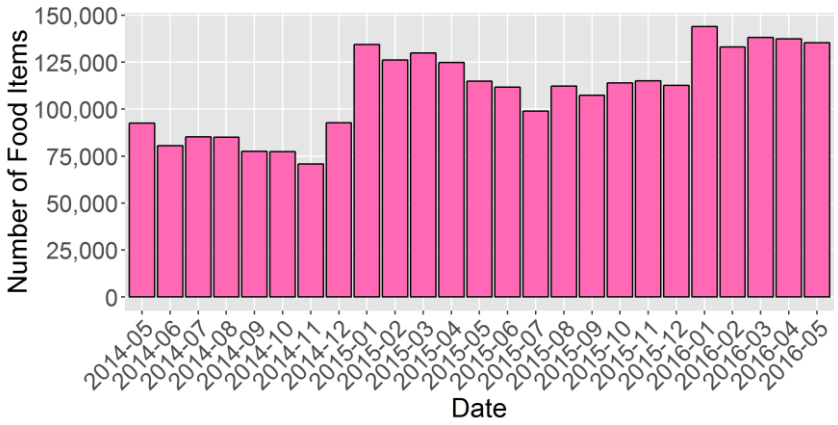

Figure 3-11: Number of food items logged in the calendar month.

3.1.6 Medications

Medications are typically recorded during a member's first coach meeting. This table includes either the medication name or the purpose of the medication. The most common medications were grouped into 14 categories. The distribution of the medication groups is shown in Figure 3-12. Blood pressure medications and antidepressants each represent 15% of all listed medications in this table. Further analysis of these groups is examined in a later section when combined with other data.

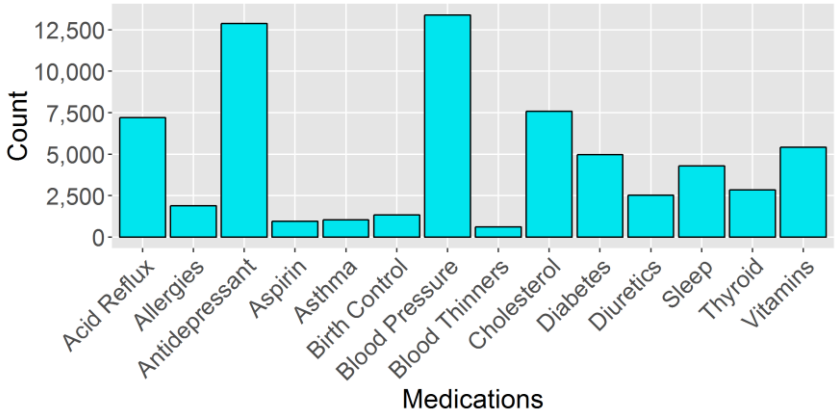

Figure 3-12: Distribution of known medication use.

3.1.7 Meal Plans

Meal plans are chosen by the weight loss coach based on the member's goals and current health. There are specific meal plans for teenagers or pregnant women whose health goals may deviate from losing weight. Most meal plans will follow Profile's three phase system: Reduce, Adapt, and Sustain.

Members typically start on a Reduce meal plan. This phase is designed to reduce food consumption in the healthiest way. When members near their weight loss goal they move to the next phase, Adapt, and start with a corresponding meal plan. Adapt meal plans are intended to transition the member into preparing their own meals while still being conscious of healthy habits formed in the Reduce phase of the program. Adapt includes a slight increase in calorie intake and an increase in activity level. Finally, members will move into phase three and begin a Sustain meal plan in which they are trying to maintain the weight loss they have achieved.

The Balance meal plan is another option. Balance is utilized for members that have special medical conditions such as members with type II diabetes on insulin, milk allergies, or someone undergoing cancer treatments. Typically, members on a Balance plan will eventually transition into Adapt and Sustain phases.

There are several meal plans to meet specific goals and each meal plan can be altered according to a member's dietary needs. By looking at the raw data, as shown in Figure 3-13, the Reboot Reduce meal plan has been used most frequently, which corresponds to the Reduce phase. By examining Figure 3-14, the Balance meal plan has the highest average number of days that members utilize the plan, 250 days. Since

Balance is designed for special dietary needs this may result in the plan being utilized longer than others.

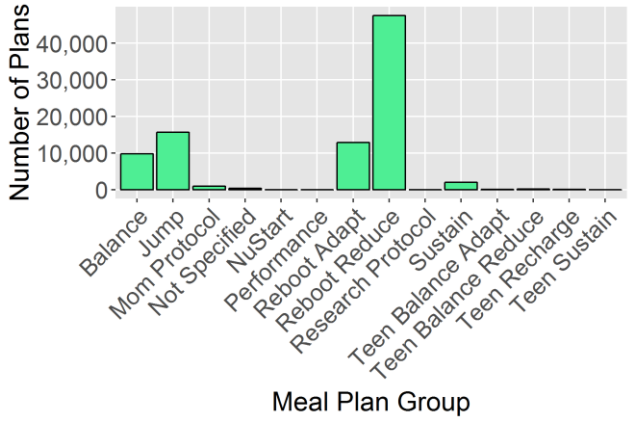

Figure 3-13: Distribution of meal plan groups.

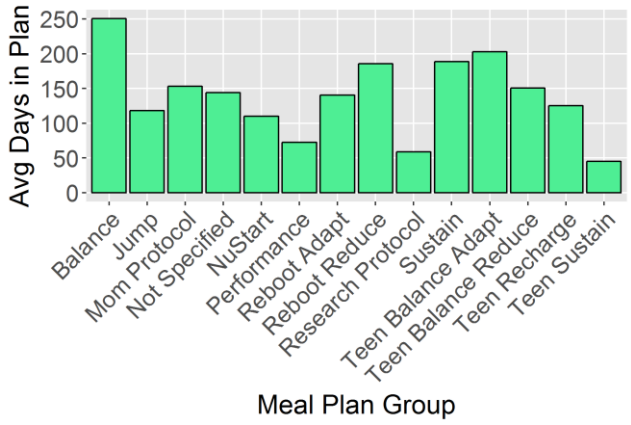

Figure 3-14: Average number of days members stay in each meal plan group.

3.2 Analysis of Combined Data Set

Each data set described in Section 3.1 is combined to represent member's monthly behavior and progress. This comprehensive data set contains one row per member per month. The data includes only those members between the age of 18 and 90 and those that are not on a pregnancy or teenager meal plan (not interested in losing weight). Excluded from the data are members that ever had an outlying weight

measurement. This data contains 365,811 observations for 33,487 unique members spanning 24 months. Although this data contains up to 24 months of data, most of the following analyses are based on 12 months of activity. Joining of this data, exclusions, and the creation of additional variables was described in detail in Section 2.1.

Figure 3-15 shows the location of each Profile member by using their zip code. Each small red dot is a Profile member and each larger blue dot is the location of a Profile store. Concentration of members appear around store locations. There are members in 48 of 50 states which includes members in both Hawaii and Alaska.

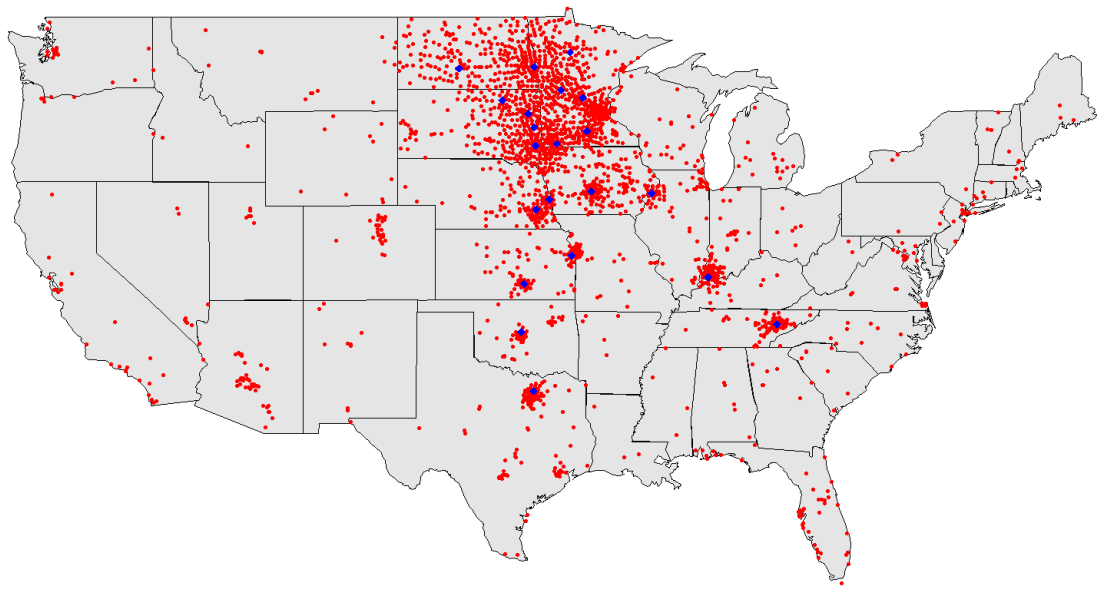

Figure 3-15: United States map depicting the location of members. Red dots represent members and blue dots represent the location of a store.

3.2.1 Distributions

Most members are females with 77.2% female and 22.3% male. This distribution is shown in Figure 3-16. When looking at the age distribution of members, we notice that most individuals (72.5%) are between the ages of 30 and 60 as illustrated in Figure 3-17. Marital status is also disclosed by each member. The pie chart in Figure 3-18 shows that

almost 59% of members are married and 33% are single. In Figure 3-19 we can see that around two-thirds of members enter the program as obese.

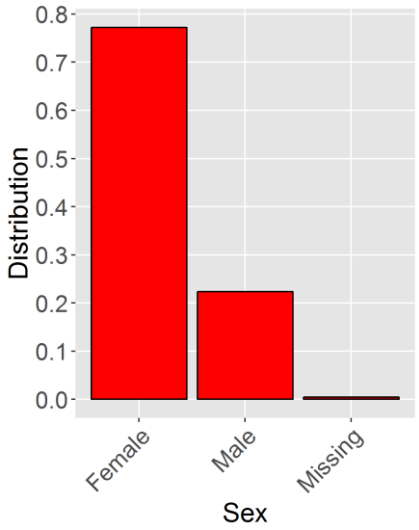

Figure 3-16: Member distribution of sex.

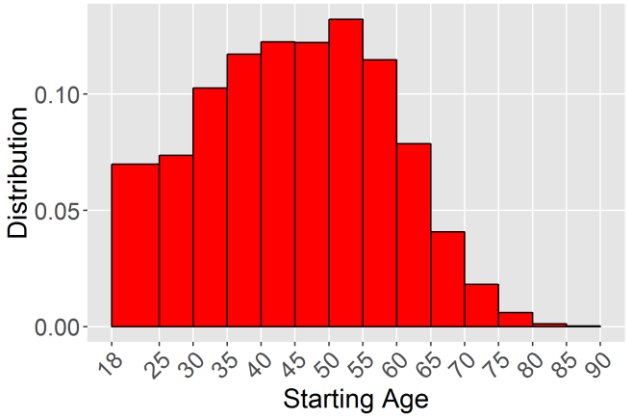

Figure 3-17: Distribution of member's age at the start of the program.

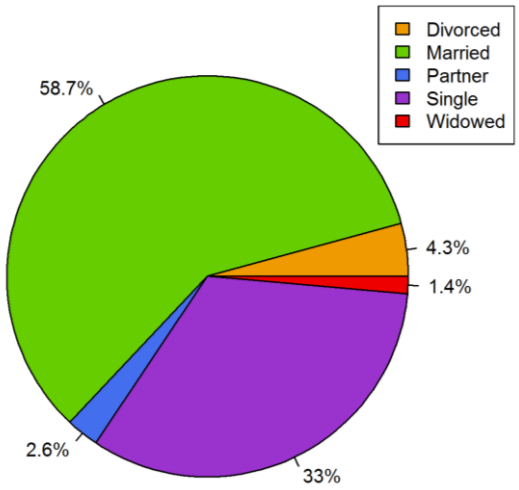

Figure 3-18: Distribution of member's marital status when starting the program.

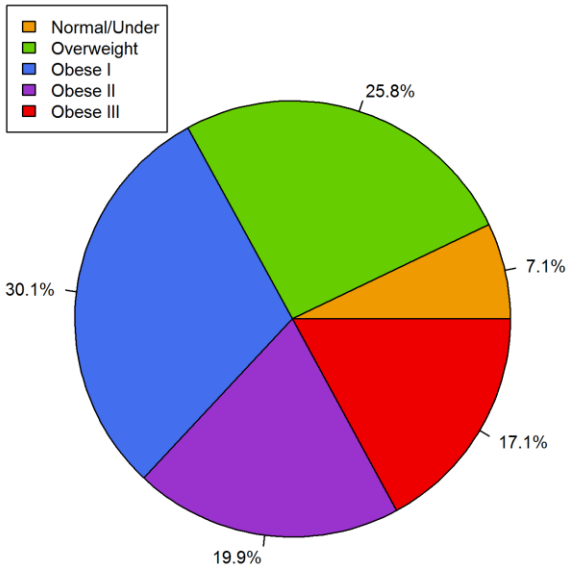

Figure 3-19: Distribution of member's starting BMI by category.

Figure 3-20 utilizes a box plot to examine starting age by sex. The average female age is 45.6 and the average male is 46.8. There does not appear to be a large age difference between males and females.

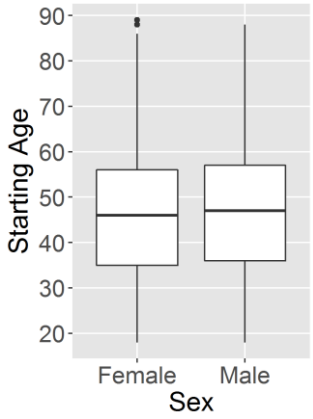

Figure 3-20: Boxplot of member's starting age by sex.

A mosaic plot, Figure 3-21, shows the distribution of members between starting BMI category and the member's sex. This plot indicates that males classified as *Normal* make up the smallest percentage of members (0.2%) whereas females classified as *Obese I* make up the largest percentage of members (22.7%).

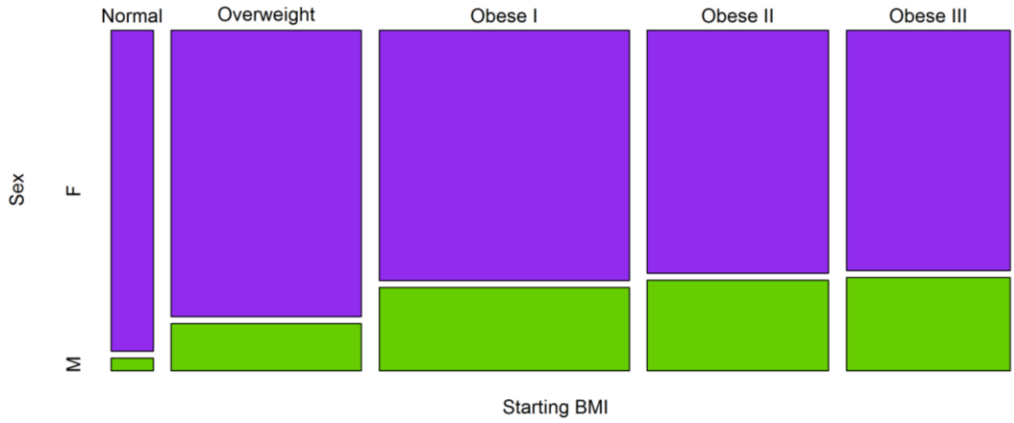

Figure 3-21: Mosaic plot of the distribution of member's sex by starting BMI category.

Figure 3-22 examines starting BMI between sex. The average starting BMI for females is 34.1 and 36.4 for males. A t-test was performed to determine if there was a significant difference between these two means at a 0.05 level. The test concluded a

significant difference in starting BMI between sex. Factors that may influence this relationship are examined in Chapter 4.

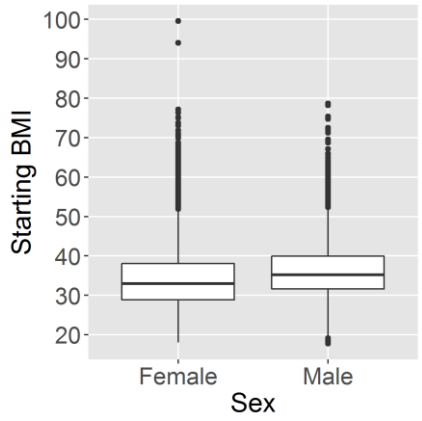

Figure 3-22: Boxplot of starting BMI by sex.

3.2.2 Medication

Medications were consolidated into 14 groups as described in Section 2.1.2.6. Figure 3-23 shows the distribution of sex within each medication group. The blood pressure medication group has a higher proportion of males than any other medication group. As a reminder, the distribution of sex in the data is 22% male and 77% female; the blood pressure medication group contains 32% male and 68% female. In contrast to Figure 3-23, Figure 3-24 shows the distribution of medication groups by females and males separately. The most noticeable difference in groups is the blood pressure group and antidepressant group. 27.4% of males are on blood pressure medication compared to 16.3% of females. This is compared to 8.9% of males and 19.9% of females on antidepressants.

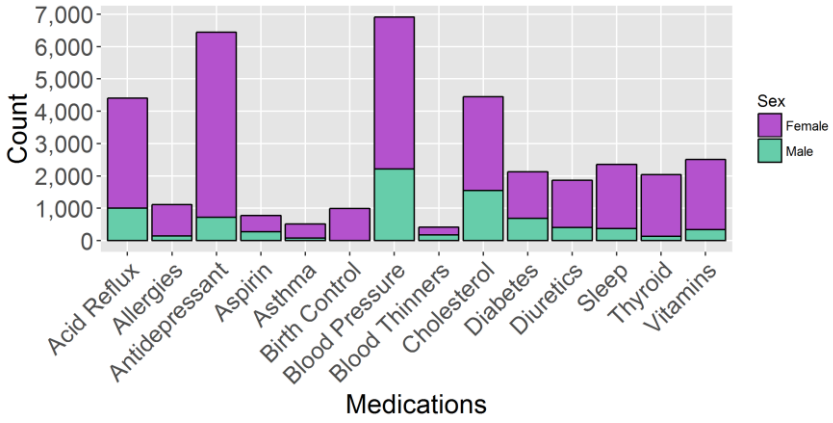

Figure 3-23: Distribution of sex within each medication group.

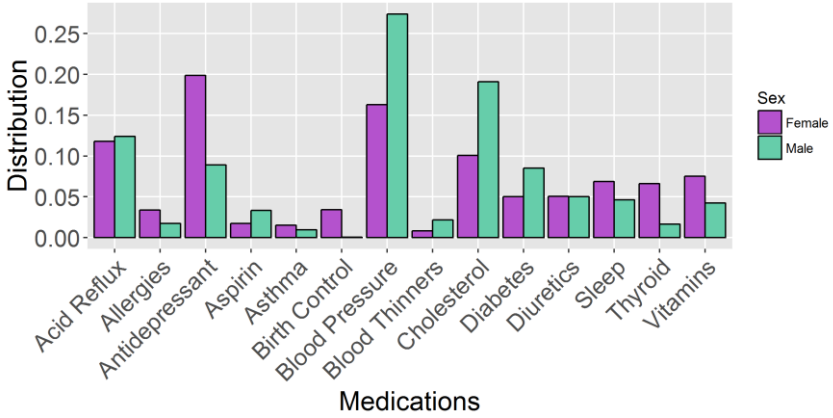

Figure 3-24: Distribution of medications by both females and males separately.

3.2.3 Weight Loss

A percentage of weight lost is calculated each month. This percentage is calculated by dividing the cumulative pounds lost since the beginning of the program by the starting weight. Figure 3-25 shows the average percentage of cumulative weight loss over time. The largest rate of decrease is within the first month in the program. The rate of weight loss decreases as time goes on and eventually turns into weight gain. By month 12, on average, members have lost 10.8% of their body weight. The shape of the curve in Figure 3-25 could also be caused by the fact that every member does not have a weight measurement at each time point. It is possible that members with high weight loss in

months 1 through 8 are not recording a weight in the later months. This behavior will be examined further in Chapter 4.

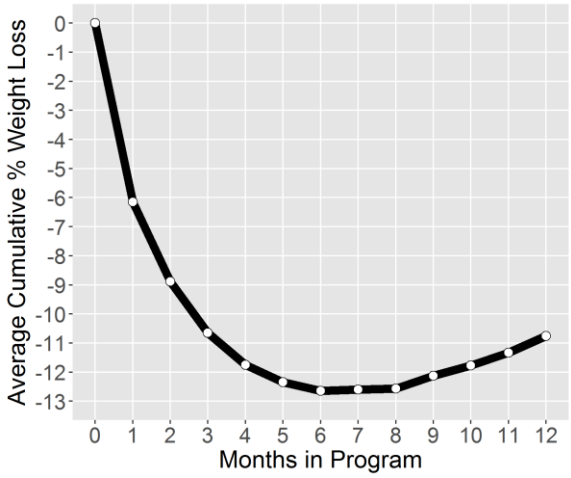

Figure 3-25: Average cumulative percentage of weight loss by each month in the program.

The average cumulative percentage of weight loss by sex is shown in Figure 3-26. Males are losing a higher percentage of weight than females in the beginning, but by month 10, the weight loss percentage is the same for both men and women. A t-test was performed for each time point and the results confirm that for months 9 through 12 the cumulative percentage of weight loss between males and females is not significantly different. This observation and test does not consider other factors that may influence this relationship.

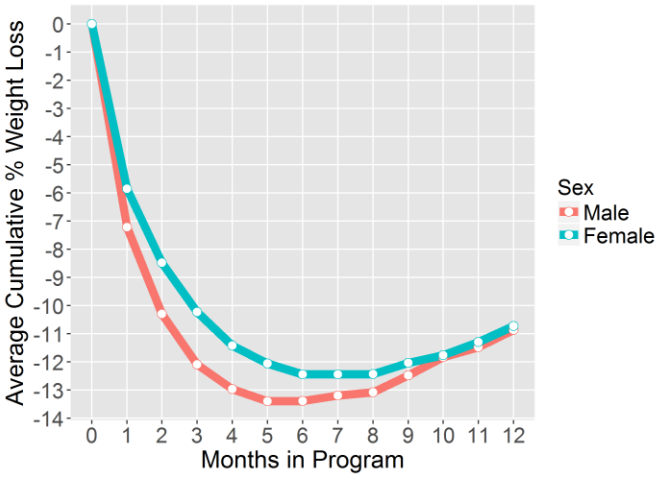

Figure 3-26: Average cumulative percentage of weight loss by each month in the program by sex.

Figure 3-27 shows the difference in cumulative weight loss percentage if a member claimed to be on any medications at the beginning of the program. We start to see a separation around month 5 where medication users are losing more of their body weight. Antidepressant use had an opposite outcome than medication use overall. As shown in Figure 3-28, if taking antidepressant medication, weight loss percentage is lower than those that are not. This observation does not consider other factors that may influence this relationship.

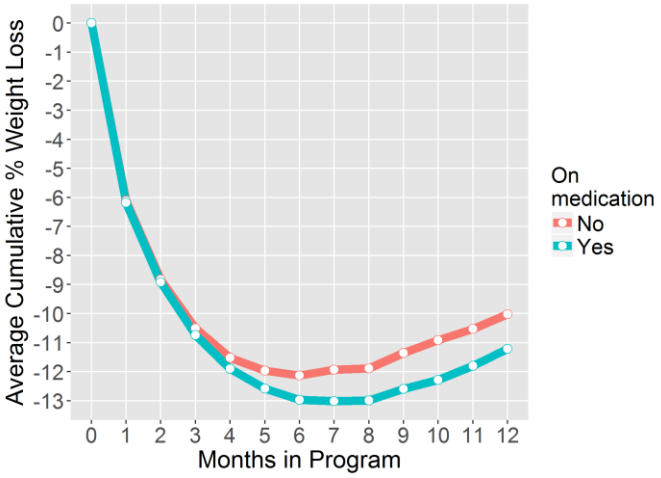

Figure 3-27: Average cumulative percentage of weight loss by each month in the program split by whether the member claims to be on medication or not.

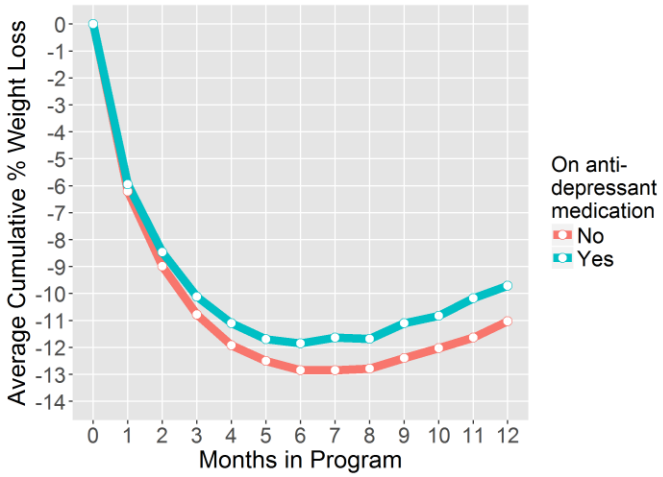

Figure 3-28: Average cumulative percentage of weight loss by each month in the program split by whether the member claims to be on antidepressant medication.

Figure 3-29 shows the average cumulative percentage of weight loss for all medication groups. Visually, there is a separation between groups; the top group (less weight loss) contains antidepressants, asthma, birth control, diabetes, no medication use, and sleep medication; the bottom group (more weight loss) contains acid reflux, allergies, aspirin, blood pressure, blood thinners, cholesterol, diuretics, any medication, thyroid, and vitamins. Removing some of the groups that contain a small number of members, Figure 3-30 shows a clearer separation. The medication groups that formed on the top (less weight loss) include antidepressants, diabetes, sleep, and none (no medications). The medication groups that formed on the bottom (more weight loss) include cholesterol, acid reflux, blood pressure, and vitamin use. Further analysis of the relationship between some of these medications and weight loss will be discussed in Chapter 4.

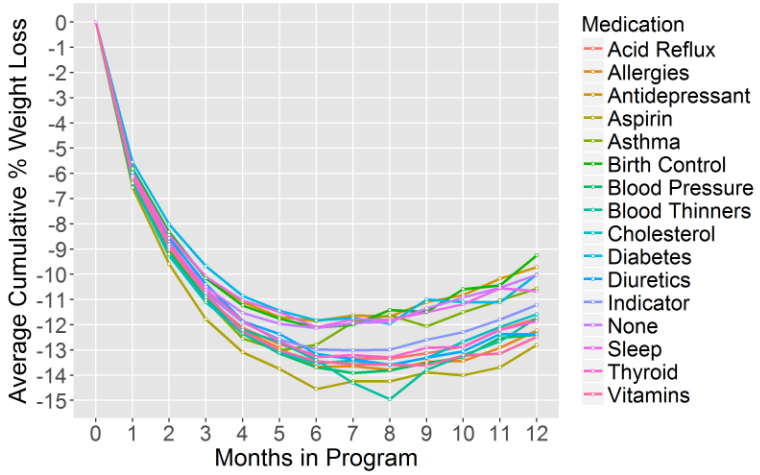

Figure 3-29: Average cumulative percentage of weight loss by each month in the program split by all medication groups.

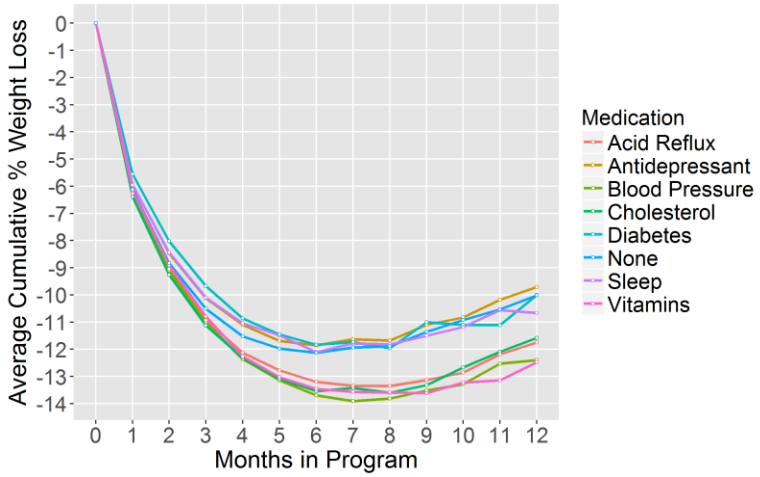

Figure 3-30: Average cumulative percentage of weight loss by each month in the program split by select medication groups.

Figure 3-31, displays the average cumulative weight loss percentage by the four most used meal plan groups: Reduce, Adapt, Sustain, and Balance. Reduce, Adapt, and Sustain meal plans appear to have similar weight loss percentages in the first three months. After month 3, Sustain continues to decrease (higher weight loss) and Reduce starts to increase (weight regain). This pattern could be attributed to the movement of members from one phase to the next. Meal plan changes and monthly representation of

these meal plans is described in Section 2.1.2.5. If a member is succeeding in the Reduce phase they will move to Adapt, then eventually Sustain. The increased weight loss in the Sustain group could be explained by successful members moving into that group while those that are not succeeding are staying in the beginning phases. The Balance meal plan shows a steady decrease in percentage weight loss through month 6 then reaches a plateau at 10% cumulative weight loss.

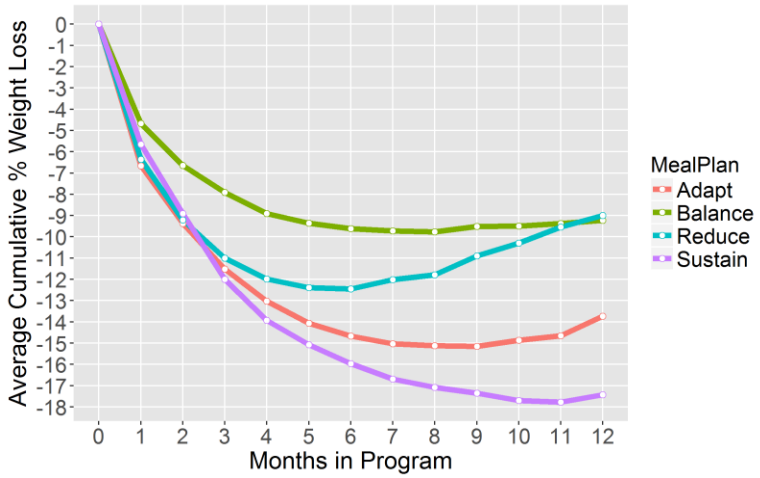

Figure 3-31: Average cumulative percentage of weight loss by each month in the program split meal plan.

3.2.4 Coach Meetings

Meeting with a weight loss coach is encouraged but not required. Figure 3-32 examines the attendance behavior of members over time. The first month in the weight loss program, 19% of members attend no meetings, 17% attend one meeting, 22% attend two meetings, 22% attend three meetings, 16% attend four meetings, and 4% attend five or more meetings. After this initial month, attendance of coach meetings decreases. The monthly distribution of no coach meetings continues to increase to 84% of members at month 12, leaving only 16% of members attending at least one meeting. Figure 3-33

shows the average cumulative coach meetings over time by sex. It appears that females are attending more meetings with their coaches than males.

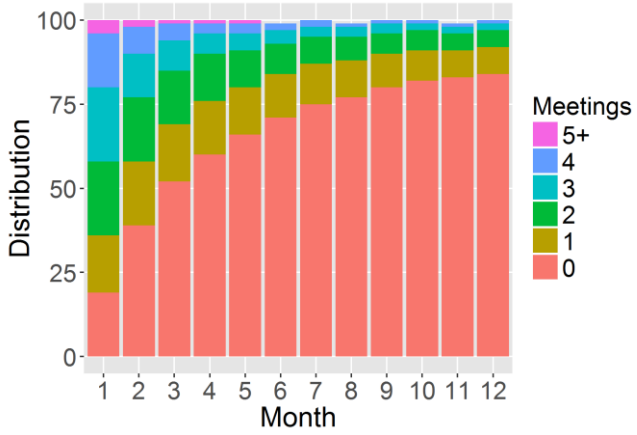

Figure 3-32: Distribution of coach meetings by the month in the program.

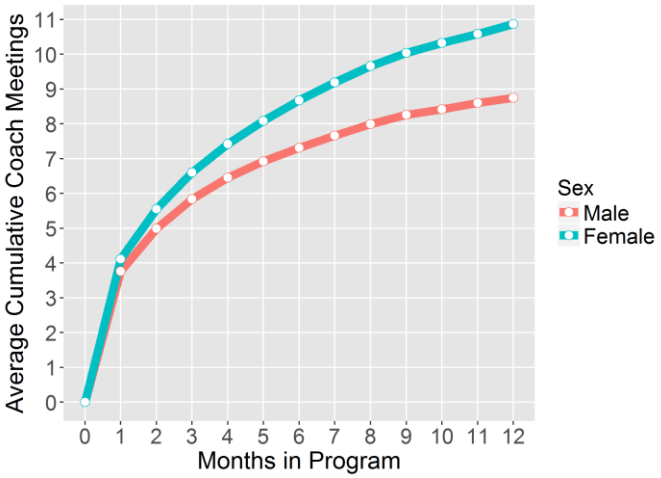

Figure 3-33: Average cumulative coach meetings by each month in the program by sex.

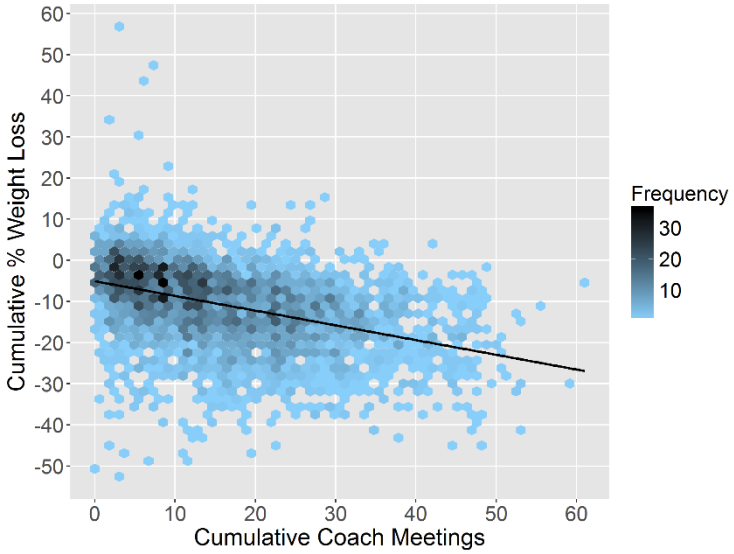

Figure 3-34: Scatterplot of member's cumulative percentage of weight loss at month 12 by the cumulative number of coach meetings at month 12. Darker colors represent a higher concentration of members. The overlaid linear regression line representing the relationship.

Examining cumulative weight loss percentage by cumulative coach meetings at month 12 in Figure 3-34, we see a relationship between weight loss and coach meetings. The graph represents any members that had a weight measurement at month 12. The darker colors represent a higher concentration of members. It is shown by the superimposed regression line that weight loss by month 12 is greater with more cumulative coach meetings within this same time.

3.3 CONCLUSIONS

Profile by Sanford has an abundance of data on each Profile member. Information ranges from how many times a day they weigh themselves to what kind of medication they take. Combining this information and getting a better understanding of customers is beneficial to Profile.

- Members weigh themselves 7 to 8 times a month.

- Most weight measurements are recorded in the morning with 68% between 5 AM and 9 AM.
- Profile members that choose to attend coach meetings are, on average, attending two meetings a month.
- The average waist-to-hip ratio for members that have these measurements is 0.88.
- Only 20% of members have utilized Profile to record their exercise and 57% have logged food items.
- 15.3% of medications used by Profile members are blood pressure medications and 14.7% of medications are antidepressant medications.
- Members that utilize the Balance meal plan average 250 days on the plan which is the longest of any meal plan.
- There are Profile members in 48 of the 50 states.
- The Profile weight loss program consists of 77% female members.
- Two-thirds of Profile members start the program as obese.
- On average, members that have recorded a weight in month 12 have lost 10.8% of their body weight.
- In their first month, 81% of members attend at least one coach meeting and by month 12, only 16% of members attend at least one coach meeting.

4 WEIGHT LOSS AT MONTH 12

Profile by Sanford utilizes one-on-one interactions between weight loss coaches and their members to encourage weight loss. Anastasiou *et al.* concluded that more interaction with a weight loss coach resulted in greater weight loss [16]. The following analyses will examine the relationship between coach interactions and weight loss at month 12. Members included in the analysis have had weight measurements all 12 months so that cumulative weight loss is accurately portrayed. This chapter will also focus on other characteristics that may influence weight loss.

4.1 DATA

The data is structured the same as described in Section 2.1 but includes only those that had at least one weight measurement per month for their first 12 months. This results in only 2262 members. A few exclusions are also applied. There is one member that has an unspecified sex, which is removed. Removed from the data is one member that has a missing age as well as 26 with a missing marital status. Since the measure of interest is percentage of weight loss, it is pertinent to exclude members that may not be interested in losing weight. Excluded from the data are any members that start with a BMI classification of underweight or normal (BMI<25) which include 124 members. Final exclusions involve certain meal plans, since four meal plans are utilized the most, those four are included in this data. Meal plans that are contained in the Adapt, Reduce, Sustain, and Balance phases of Profile's weight loss program are only included, excluding 28 members. These exclusion criteria apply to some members more than once so we are left with 2087 members.

4.2 VARIABLES

The outcome variable is the cumulative percentage of weight lost. The percentage of weight loss over time for this group of members is shown in Figure 4-1. Covariates taken into consideration for the following analyses are analogous to those discussed in previous chapters; sex, age, marital status, starting BMI, medication use, and coach meeting attendance.

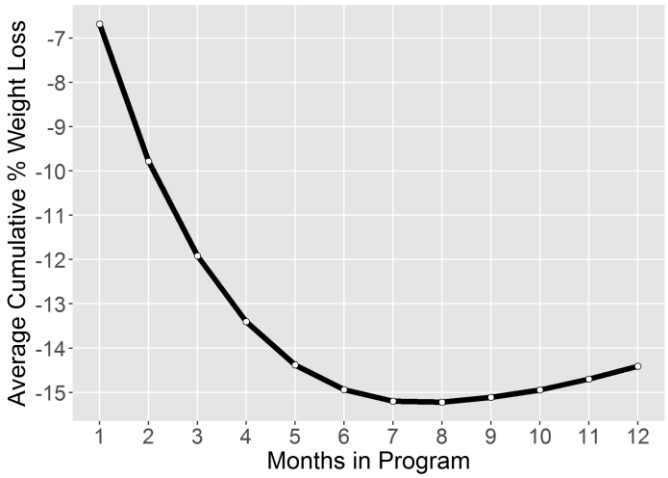

Figure 4-1: Average cumulative percentage of weight loss by each month in the program.

Distributions of nominal and binary variables are examined in Table 4-1. For clarification, 66% of the 2087 members use at least one medication. The remaining medication percentages represent the distribution of members that utilize that medication. With this group of members, we can see that blood pressure and antidepressant medications are highly used, 27% and 20%, respectively. According to the National Health and Nutrition Examination Survey (NHANES), which is a cross-sectional survey of noninstitutionalized Americans, in 2011 and 2012, 13% of adults (20 years and older) took antidepressants and 5.4% took medication for high blood pressure [39]. These

numbers are similar to the CDC report for 2009-2012 which states 9% of Americans (all ages) take antidepressants and 4.3% take blood pressure medication [40].

Table 4-1: Distribution of nominal and binary variables.

	Month 12 (N = 2087)
Sex	
Female	79.0%
Male	21.0%
Marital Status	
Relationship	61.2%
Single	38.8%
Medication Use	66%
Antidepressant	20%
Blood Pressure	27%
Cholesterol	20%
Diuretics	8%
Diabetic	8%
Sleep	8%
Acid Reflux	17%
Vitamins	11%
Allergies	5%
Aspirin	3%
Blood Thinners	2%
Thyroid	8%

Table 4-2 describes the continuous variables included in the data. This table displays each variable mean and the standard error of the mean along with the minimum and maximum values. Negative values for variables such as "Weight loss, %" denote weight loss whereas positive values denote weight gain. The average age of the members actively engaged in the weight loss program is 50 years with starting weight of 221 pounds. After 12 months in the program these members weigh an average of 187 pounds (the same average weight at month 6) with a range of weight loss from almost 200 pounds lost to 30 pounds gained. The independent variables in Table 4-1 and Table 4-2 are baseline measurements unless specified as being measured at month 6 or month 12.

Body weight, BMI, weight loss %, weight loss pounds, and cumulative coach meetings are measured at months 6 and 12.

Table 4-2: Summary statistics on continuous variables.

	Mean ± SE	Range
Age, years	50.68 ± 0.26	20 - 90
Height, inches	66.17 ± 0.08	56 - 80
Starting weight, pounds	221.00 ± 1.13	134 - 555
Starting BMI	35.34 ± 0.15	25 - 75
Measured at Month 6		
Body weight, pounds	187.14 ± 0.94	103 - 380
BMI, kg/m²	29.93 ± 0.13	20 - 55
Weight loss, %	-14.94 ± 0.16	-36 - 11
Weight loss, pounds	-33.86 ± 0.45	-175 - 31
Cumulative Coach Meetings	13.83 ± 0.15	0 - 42
Measured at Month 12		
Body weight, pounds	187.83 ± 0.94	104 - 382
BMI, kg/m²	30.03 ± 0.13	20 - 56
Weight loss, %	-14.41 ± 0.21	-48 - 15
Weight loss, pounds	-33.17 ± 0.58	-199 - 30
Cumulative Coach Meetings	20.35 ± 0.26	0 - 61

Spearman's correlation was calculated for each pair of variables to examine the relationship between cumulative percentage of weight loss and the variable. Table 4-3 displays Spearman's correlation and the corresponding p-value for each variable with the cumulative percentage of weight loss at month 12. Each medication variable represents the use of that variable, therefore if a member uses that medication the value is 1, otherwise 0. Some variables may be missing since they were not significantly correlated (p-value < 0.05) with percentage of weight loss. Variables such as antidepressant, diabetes, and vitamins as well as starting weight and starting BMI are significant (p-value < 0.05). Additionally, body weight, BMI, weight loss percentage, weight loss in pounds, and cumulative coach meetings by month 6 are also significantly correlated with the cumulative percentage of weight loss by month 12. The total number of coach meetings

attended through month 12 is significantly correlated with the percentage of weight lost by that time.

Table 4-3: Spearman's correlation coefficient and p-value with the cumulative percentage of weight loss at month 12.

	Percentage of Weight Loss	
	Spearman's Rho	p-value
Age, years	-0.0358	0.1023
Sex (Male=1, Female=0)	0.0353	0.1071
Medication Use	-0.0277	0.2052
Antidepressant	0.0697	0.0014
Blood Pressure	-0.0710	0.0012
Diuretics	-0.0269	0.2186
Diabetic	0.0627	0.0042
Vitamins	-0.0629	0.0041
Height, inches	0.0223	0.3077
Starting weight, pounds	-0.2713	0.0000
Starting BMI	-0.3315	0.0000
Measured at Month 6		
Body weight, pounds	0.0710	0.0012
BMI, kg/m²	0.0718	0.0010
Weight loss, %	0.8875	0.0000
Weight loss, pounds	0.8324	0.0000
Cumulative Coach Meetings	-0.3166	0.0000
Measured at Month 12		
Cumulative Coach Meetings	-0.3310	0.0000

4.3 RESULTS

Figure 4-2 represents the cumulative percentage of weight loss at month 12 by the cumulative number of coach meetings. The darker dots represent a higher concentration of members. The lighter line represents the linear regression line, the black line represents the Loess curve, and the gray line represents a cubic polynomial. It is evident that a relationship exists between coach meetings and weight loss. All three fitted lines suggest a positive relationship between a higher weight loss percentage and attendance of more coach meetings.

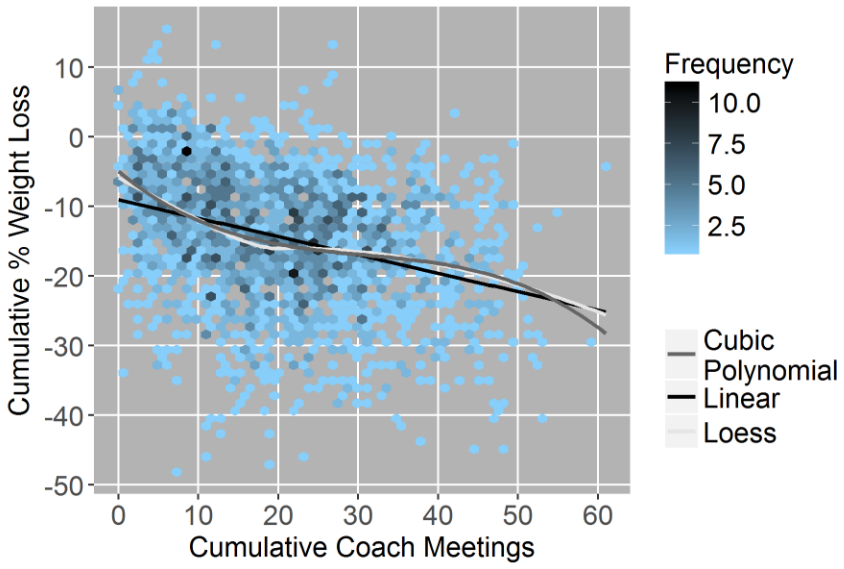

Figure 4-2: Cumulative percentage of weight loss at month 12 by cumulative number of coach meetings at month 12. Darker areas denote more observations. The light line is the linear relationship between weight loss and coach meetings. The black line represents the Loess curve to fit a smooth relationship between the two variables. The gray line represents a cubic polynomial.

The linear regression line in Figure 4-2 is described in Table 4-4 by a regression model. After 12 months in the program and no coach interaction, members still lose an average of 9% of their body weight. With each additional cumulative coach meeting, members lose 0.26 percentage points more of their body weight. If a member attends, on average, one more coach meeting a month (over 12 months), they lose 3.2 percentage points more of their body weight.

Table 4-4: Linear regression model results for cumulative percentage weight loss at month 12 as the dependent variable.

Variable	Estimate	P-Value
Intercept	-9.0404	< 0.0001
Cumulative Meetings	-0.2639	< 0.0001

Monthly coach meetings and percentage of weight loss could be influenced by a confounding factor. Confounding factors are sometimes overlooked in observational

studies [41]. As an example, if women tend to lose more weight than men but women also attend more coach meetings than men, sex could be a confounding factor. Since sex influences the number of coach meetings and the weight loss percentage, it is possible that coach meetings are not as influential on weight loss but a result of a member being male or female. One definition of a confounding factor or confounder given by Vanderweele and Shpitser states that if a parameter estimate stays the same after adjusting for the covariate, then it is not a confounder [42]. A confounder is expected to change the estimate by more than 10 percent [42]. Based on the idea of confounding factors, several covariates were examined on their effect on coach meetings and cumulative weight loss. Potential confounding factors were sex, age, marital status, starting weight, starting BMI, total medications, medication indicator, and use of specific medications (antidepressants, vitamins, diuretics, blood pressure, diabetic and sleep medications). Starting weight and starting BMI changed the parameter estimate of the number of coach meetings by more than the 10% threshold. Table 4-5 shows the regression model when adding starting BMI.

Table 4-5: Regression model on cumulative percentage of weight loss by month 12 with coach meetings and starting BMI as covariates.

Variable	Estimate	P-Value
Intercept	2.7139	0.0055
Cumulative Meetings	-0.2122	< 0.0001
Starting BMI	-0.3624	< 0.0001

The parameter estimate changed from -0.2639 to -0.2122 after adding starting BMI to the model. When holding starting BMI constant, each increase in the number of cumulative coach meetings results in an average 0.21 percentage point increase in cumulative weight loss. Now, accounting for starting BMI, member lose, on average, 2.5

percentage points more of their body weight with one more coach meeting a month. Controlling for starting BMI decreases the portrayed effect of coach meetings but the effect is still significant. Sex, age, marital status, number of medications, or specific medication use for high blood pressure, diabetes, antidepressants, vitamin use, diuretics, and sleep medications were not found to be confounding factors between weight loss and coach meetings.

Other covariates appear to influence cumulative percentage of weight loss but confounding factors need to be considered. A linear regression on cumulative weight loss at month 12 by whether the member takes blood pressure medication appears to suggest that taking blood pressure medication influences weight loss as displayed in Table 4-6. There are several classes of medications to control high blood pressure, each with their own side effects which include both weight gain and weight loss [43,44]. The parameter estimate for blood pressure medication use changes after adding starting BMI into the regression model. When controlling for starting BMI, blood pressure medication use is not significant to weight loss but instead starting BMI is influencing that relationship.

Table 4-6: Regression model on cumulative percentage of weight loss by month 12 with blood pressure medication and starting BMI as covariates.

Variable	Without Starting BMI		Include Starting BMI	
	Estimate	P-Value	Estimate	P-Value
Intercept	-14.0018	< 0.0001	1.4681	0.1470
Blood Pressure Mediation Use	-1.5084	0.0012	-0.2385	0.5930
Starting BMI	--	--	-0.4475	< 0.0001

Weight gain is a side effect of some antidepressants [45]. The relationship between depression and obesity has also been studied and suggests an association

between the two especially among women [45]. Table 4-7 shows the linear regression results of two different models. Including antidepressant use as the only covariate in the model, then adding starting BMI and sex to the model. Interaction between antidepressant use and sex was also examined but was found to be not significant in the model (p-value > 0.05).

Table 4-7: Regression model on cumulative percentage of weight loss by month 12 with antidepressant medication, starting BMI, and sex as covariates.

Variable	Only Medication		Include BMI and Sex	
	Estimate	P-Value	Estimate	P-Value
Intercept	-14.7421	< 0.0001	1.2540	0.2120
Antidepressant Use	1.6377	0.0014	2.3442	< 0.0001
Starting BMI	--	--	-0.4682	< 0.0001
Sex (1=Male, 0=Female)	--	--	1.9228	0.0001

After controlling for starting BMI and sex, the estimated effects of antidepressant use on weight loss changes. With both starting BMI and sex held constant, on average members are losing 2.3 percentage points less when taking antidepressants. A difference in cumulative percentage of weight loss by month 12 exists between male and females for this subset of Profile members as shown by the model in Table 4-7. Male cumulative percentage of weight loss is 1.9 percentage points less than females when starting BMI and antidepressant use is held constant.

4.4 CONCLUSIONS

Profile by Sanford focuses much of its weight loss program on utilizing motivation and support given by weight loss coaches. When measuring weight loss of members that actively participate in the program, more coach meetings is associated with an increased weight loss. Increasing a member's coach meeting attendance to one more

meeting a month results in, on average, 2.5 percentage points more weight loss for Profile members who weigh themselves consistently each month for the first 12 months in the program.

Other factors appear to influence weight loss such as taking blood pressure medication. After controlling for starting BMI, this medication is no longer a significant factor in weight loss. Blood pressure medication does not result in weight loss but rather, those with higher BMI are taking this medication and those with higher BMI are losing more weight. Medications such as antidepressants are associated with weight gain where this group of Profile members are seeing 2.3 percentage points less weight loss if taking antidepressants. Additionally, females are losing 1.9 percentage points more than males for this subset of Profile members. Therefore, antidepressant use, starting BMI, and sex are all associated with cumulative percentage of weight loss at month 12 for Profile members that have weight measurements each month.

5 JOINT MODELING FOR TIME TO DROPPING OUT OF THE PROGRAM

Time until an event occurs is a question that presents itself in several areas of research. Examples include the time until the onset of an illness in medical research, machine failure in industrial research, or loan default in financial research. Inquiries may rise about the relationship between the time of these events and a repeated measurement. For example, a repeatedly measured biomarker in a patient and its association with the time until the patient's death or illness. Joint modeling is a method that measures this relationship by combining the time-to-event analysis with a longitudinal model on the repeatedly measured characteristics, both depending on a standard set of random effects.

5.1 LONGITUDINAL MODEL

In predictive modeling, the terms *repeated measures* and *longitudinal* are used interchangeably [31]. A repeated measures study is defined as an individual (or any other unit) that is observed at two or more times or places throughout the study [31]. Data collected over time is typically referred to as longitudinal data [31]. Longitudinal data analysis is useful when the possibility of correlation between observations on the same individual arise [31]. Since independence between observations is an assumption in simple regression models, techniques such as linear regression are not appropriate for longitudinal data. Data in which individuals have multiple measurements of a covariate over time can be modeled using a linear mixed effects model. The mixed model was run using the nlme package [46] in R.

5.1.1 Data

The data described in Chapter 3 is the starting point for the data utilized throughout Chapter 5. The data include only members that are age 18 to 90. To remove outliers due to measurement errors, weight loss measurements that had a large deviation from previous measurements were also excluded. Members not actively trying to lose weight are excluded by looking at a member's meal plan. Anyone that has been on a "Teen" or a "Mom Protocol" meal plan is excluded from the data. Only members that are classified as overweight or obese with BMI of 25 or higher were included.

Data used to build the mixed model is in longitudinal form. Each member of Profile has an observation for each month after they joined the program through month 12. Member's starting Profile in May 2014 through April 2015 are included in the data. Information is recorded for these individuals through April 2016. Someone starting the program in May 2014 could have 12 months of data recorded through May 2015. This structure allows all members to have reached 12 months in the program.

The data was divided into two data sets, one for developing the model and one for validation. Members were randomly assigned with 70% in the development data and 30% in the validation data. Variable selection and model development was done using the development data and model selection was done by applying the model to the validation data. Since the development and validation data was only retrieved through April 2016, results in May 2016 were used as hold-out data. This hold-out sample allows for assurance that the model is performing as expected for further validation. The full data consists of 10,022 different Profile members with 60,125 total observations. The

development data contains 7015 members with 42,210 observations and validation data contains 3007 members with 17,915 rows of data.

5.1.2 Response Variable and Covariates

Monthly percentage of weight loss is the outcome in the mixed model,

$$y_{i,j} = \left(\frac{weight_{i,j} - weight_{i,j-1}}{weight_{i,j-1}}\right) * 100, \qquad 5\text{-}1$$

where i represents a Profile member and j represents time in months. This measurement is the percentage of weight lost since the preceding month. A negative $y_{i,j}$ represents weight loss while a positive $y_{i,j}$ represents weight gain. The first covariate to consider in the model is time. As shown in Figure 5-1, monthly percentage of weight loss over time is not linear so we cannot expect a linear model to fit the data. Spline curves are piecewise polynomial curves used to fit non-linear data. Spline functions can be generated to fit these curves. Spline functions can be complex depending on the degree of the function and the number of internal knots as described in Section 2.2.1.

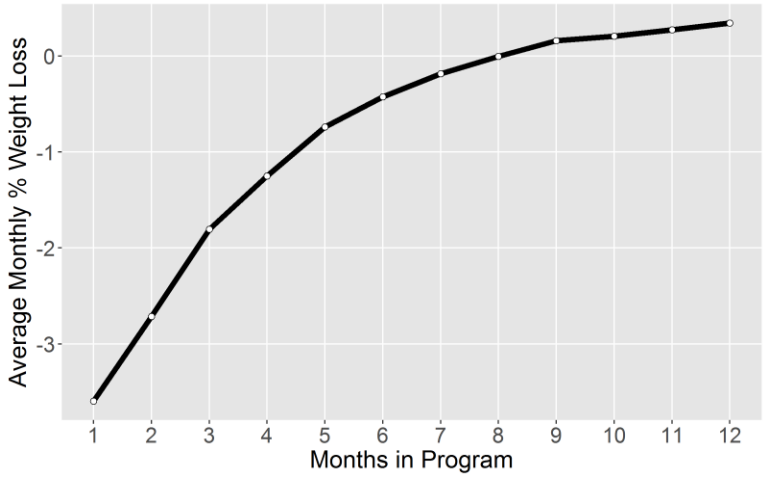

Figure 5-1: Average monthly percentage of weight loss by month in the program.

Several spline functions were generated with varying degrees and knot locations. A spline function of degree 1 did not fit the curve as nicely as degree 2 and degree 3 functions. Since a third-degree spline did not show a large improvement in performance from a second-degree spline, a spline function of degree 2 was utilized. Knot locations is also important for these functions. A spline function of degree 2 with no knots was examined then additional knots were added based on the shape of the curve in Figure 5-1. Adding knots to the function increased the predictability and complexity of the mixed model. Maximizing the log-likelihood function, as mentioned in Section 2.2.4.2, is computationally extensive. An increasingly complex mixed model, such as utilizing a high degree spline function with several knots plus additional variables, can add to the computation time. For this reason, a second-degree spline function with knots at time 3 and 6 was selected before adding more covariates to the model.

Variables that were considered in the mixed model are listed in Table 5-1. Each of the monthly covariates were lagged one month. This is necessary if the model is used for predictions. For example, if a member is starting month 6 in the program and we want to apply the mixed model, we would only have information recorded through month 5. If the number of coach meetings was a covariate in the model, we would have this measurement for months 1 through 5. Therefore, each of the monthly calculated covariates considered in the model are lagged one month.

Table 5-1: Variables considered in the mixed model.

Variables		
Age	Indicator of medication use	Number of activities recorded*
Sex	Meal plan*	Number of food items logged*
Marital Status	Weight*	Number of coach meetings*
Starting BMI	BMI*	Cumulative number of coach meetings*
First weight	Weight loss*	Cumulative percentage of weight loss*
Total medications	Number of weight recordings*	Monthly percentage of weight loss*

*measured in previous month

Variable selection was done to determine the best potential combination of covariates for the model. It was found that actual weight and BMI were correlated as well as cumulative percentage of weight loss and actual pounds lost. When considering variables in the model, only one from the correlated pair was utilized. Pearson's correlation coefficient was examined between each covariate and the response variable at each month. Covariates were ranked according to the absolute value of the correlation coefficient. Each of the 12 ranks were averaged to obtain an overall ranking of the covariates.

Each covariate was considered in a mixed model with the second-degree basis spline function with knots at time 3 and 6. The root means square error (RMSE) and the model's Akaike Information Criterion (AIC) were examined. These two measures ordered variables to determine importance in predicting monthly percentage of weight loss. Correlation with the dependent variable, RMSE, and AIC were all considered and variables were ranked according to these three measurements. The top five variables are shown in Table 5-2.

Table 5-2: Top variables for longitudinal model after variable selection process.

Variable	Overall Ranking
Coach meetings in previous month	1
Meal plan in previous month	2
Starting BMI	3
Number of weight measurements in previous month	4
Starting Weight	5

The variable selection process indicates that the number of coach meetings, the meal plan, and the number of weight measurements in the previous month as well as starting BMI weight are predictive of the current month's percentage weight loss. Starting weight and starting BMI are correlated so only starting BMI will be considered in the mixed model. Starting BMI and the other variables in Table 5-2 are considered as potential covariates in the mixed model along with sex and age.

5.1.3 Model

Time and members are considered random variables in the model [31]. We will start with the simplest model including only time as a fixed effect and both time and member as random variables. A random intercept model assumes that each member has the same slope over time but have a differing baseline weight loss,

$$y_{ij} = \beta_0 + \beta_1 t_{ij} + b_{i0} + e_{ij}, \qquad 5\text{-}2$$

where the term $(\beta_0 + b_{i0})$ represents the intercept for the i^{th} member and β_1 is a fixed effect for the average slope between members [30]. The random intercept model assumes that the correlation between weight loss percentages is constant over time. We might expect that measurements that are closer in time are more correlated than measurements that are taken farther apart.

The random slope model adds a random effects term. This model allows for members to have different slopes;

$$y_{ij} = \beta_0 + \beta_1 t_{ij} + b_{i0} + b_{i1} t_{ij} + e_{ij}. \quad 5\text{-}3$$

The term $(\beta_1 + b_{i1})$ represents the slope for the i^{th} member. Each member has their own intercept and slope in the model. We can incorporate the spline functions into the random slope model and add covariates to improve predictability. The model selection process started by generating several mixed models with a combination of covariates and examining the AIC and RMSE when applied to the validation data.

As mentioned in Section 5.1.2, the number of coach meetings is predictive of the percentage of monthly weight loss. It was found that adding additional variables to this model slightly improved the model. A log-likelihood test shows a significant improvement (at a 0.05 significance level) when adding additional variables into the model. *Model 1* contains only the number of coach meetings with the spline function of time. *Model 2* includes the spline function of time and coach meetings but adds starting BMI. *Model 3* includes the spline function of time, coach meetings, starting BMI and the number of weight measurements recorded by the end of the month. Age and sex were also considered as covariates but did not improve the model AIC or RMSE. The AIC and RMSE for the three models are compared in Table 5-3.

Table 5-3: Three mixed model descriptions, AIC, and RMSE.

Model	Description	AIC	RMSE	
			Validation	May
1	Spline(months) + Number of Coach Meetings	161670	1.8250	1.8795
2	Model 1 + Starting BMI	161528	1.8130	1.8623
3	Model 2 + Number of Weight Measurements	160937	1.8073	1.8598

The RMSE was calculated by applying each model to the validation data and the hold-out data from May 2016 (displayed as "May"). The RMSE measurement is based on the fixed effects predictions. Both AIC and RMSE show slightly smaller numbers in *model 3*. Visual representation of the models can be compared in Figure 5-2 which displays the average monthly percentage of weight loss predicted by each model when applied to the validation data compared to the actual values.

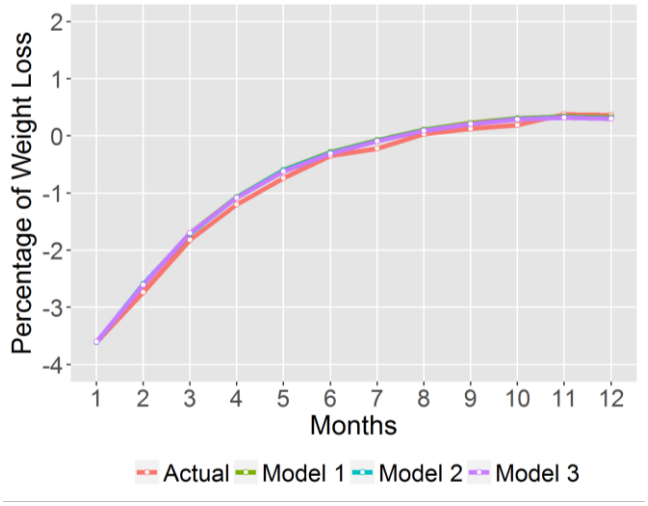

Figure 5-2: Comparison of three candidate models applied to validation data.

The average predictions for all three models are nearly on top of each other. We know that the RMSE is smallest for *model 3* but Figure 5-2 shows only a negligible difference between the models. A log-likelihood test was conducted to determine if the reduced models were better than adding additional variables. It was concluded that adding additional variables improved the model. Since statistically, *model 3* outperforms *model 1* and *model 2*, this model will be utilized as the mixed model to predict the monthly percentage of weight loss. The final output for *model 3* is shown in Table 5-4.

Table 5-4: Model output for *model 3*, the final mixed model.

Variable	Coefficient	p-value
Intercept	-3.7745	< 0.0000
Spline1(months)	2.4467	< 0.0000
Spline2(months)	4.3294	< 0.0000
Spline3(months)	5.3609	< 0.0000
Spline4(months)	5.2604	< 0.0000
Coach Meetings	-0.1281	< 0.0000
Weight Recordings	-0.0157	< 0.0000
Starting BMI	-0.0259	< 0.0000

The spline function contains four basis functions, so these coefficients are not easily interpretable. With month and all other variables held constant, for one additional coach meeting a member loses 0.13 percentage points more monthly; one more weight measurement recorded they lose 0.02 percentage points more monthly; and with a unit increase in starting BMI, on average, a member loses 0.03 percentage points more.

5.2 TIME-TO-EVENT MODEL

Survival analysis is a method used to measure the association of covariates to an event. The name *survival analysis* originated in health care when it was used to determine the time until the death of patients [47]. Since then survival analysis has expanded to other industries [47].

5.2.1 Data

The survival data is not longitudinal in structure. Instead, there is one row per member. Each row contains the member's ID, baseline covariates, time, an event indicator, and censoring information. If a member does not have a weight measurement or did not meet with a coach for an entire month, this is considered an event. For example, if a member did not have either of these measurements in their fourth month in

the program, their event indicator would be 1 and their time variable would be 4. On the contrary, if a member had reached their fourth month but recorded a weight measurement, their event indicator would be 0 and their time variable would be 4. Members were considered right-censored if they had not reached month 12 by April 2016 (when the data was pulled) or if they had reached their twelfth month and were still active.

For the data to be utilized in the future joint model, the longitudinal and survival data need to contain the same group of members [38]. Therefore, the survival data was also divided into development, validation, and hold-out data sets. The same members in the longitudinal development data set are in the survival development data set. The survival development data consist of 70% of the total members.

The longitudinal data contains a row for each month a member is active and the survival data has one row of data indicating if that member has dropped out or if they are a censored observation. For example, if a member drops out in month 5 the longitudinal data contains a row for months 1 through 4, and the survival data includes one row and indicates that month 5 was the drop-out month. Consequently, if a member drops out in month 1, the survival data indicates this in the one row of data for that member but the longitudinal data would contain one row for month 0. This is a problem since the longitudinal data does not provide any month 0 information. Theoretically, this information could be included and consists of baseline data except that it was decided to include previous month information in the longitudinal data. Therefore, the previous month to month 0 does not exist and hence there are no drop-outs in month 1 in the data.

The full data contains 10,022 rows of data with 7015 in the development data and 3007 in the validation data. A Kaplan-Meier plot is generated from the development data as mentioned in Section 2.2.3.1 and shown in Figure 5-3. The Kaplan-Meier plot shows the probability of survival at each time point of those at risk. Survival in this plot refers to members not dropping out of the program. Figure 5-3 shows that at month 1 all members are still in the prorgam, by month 6 less than half of the members are still in the program, and by month 12 only 20% have not dropped out.

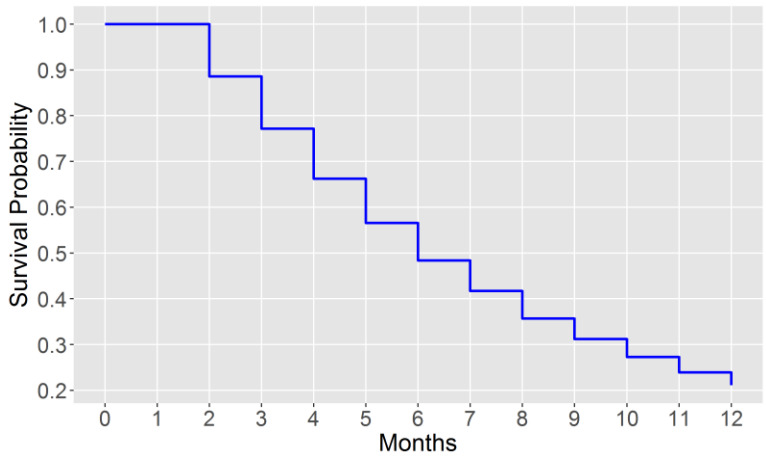

Figure 5-3: Kaplan-Meier plot.

5.2.2 Covariates

The covariates considered in the survival model are measured at baseline. Sex, age, marital status, starting weight, starting BMI and starting month were all considered. Exogenous time-dependent covariates are not considered in the model.

5.2.3 Model

Each variable mentioned in Section 5.2.2 was considered in a Cox proportional hazards model to determine the predictability of each covariate separately. The AIC of several models was examined with varying combinations of variables. The concluding

model contains age, marital status, and sex with the results shown in Table 5-5. The coefficient estimates, $\hat{\beta}$, are log hazard ratios and subsequently $e^{\hat{\beta}}$ is the hazard ratio when the corresponding covariate is increased one unit with all other covariates held constant [30]. Marital status is a categorical variable with reference value being "missing".

Table 5-5: Survival model output.

Variable	Coefficient	exp(Coefficient)	p-value
Age	-0.0135	0.9865	< 0.0001
Marital Status - Relationship	-0.2498	0.7870	0.0155
Marital Status - Single	-0.0828	0.9328	0.4830
Sex - Male	0.0553	1.0672	0.0400

With a unit increase in age, there is 1.3% decrease in the risk of an event. As age increases, members are more likely to stay in the Profile program. Members that are in a relationship (married or partner) have a 22% lower risk than members with a missing marital status (1.8% of members). Single (single, divorced, widowed) members have a 7.9% lower risk. Simply stated, when marital information is missing their risk of an event is highest, followed by single members, and those that are in a relationship are most likely to participate in the program. All these numbers are based on the idea that age and sex are held constant. Finally, the model concludes that men have a 5.7% higher risk of an event than females. Females are more likely to stay in the program than men. This model is written as,

$$h_i(t) = \exp\Big\{-0.0135 * age_i + \ -0.2498 * MS_{relationship_i} + \ -0.0828 * MS_{single_i} + \ 0.0553 * gender_{male_i}\Big\}. \quad 5\text{-}4$$

The probabilities generated from the survival model are described as the probability of "surviving" (not dropping out). When examining the accuracy of the model, a probability cutoff value of 0.84 was chosen to classify members as either 1 (drop-out) or 0 (stayed in the program). If a member's probability of not dropping out of the program is less than 0.84 they are classified as a 1 (drop-out), otherwise they are classified as a 0 (stay in the program).

Table 5-6: Survival model performance measures.

Measure	Development	Validation	May
AUC	0.5630	0.5616	0.5656
Accuracy	77.94	78.90	78.48
True Positive	2.48	2.37	2.44

Table 5-6 describes how well the model performs when applied to the development, validation, and hold-out data ("May"). Predictive models should demonstrate consistency in their predictive abilities. In Table 5-6 it is good to see that the Area Under the Curve (AUC), accuracy and true positive rates are consistent throughout the development, validation, and May data.

5.3 JOINT MODEL

Joint modeling is an enhancement of the Cox proportional hazards model. This method predicts the probability of an event in time. In addition to baseline covariates found in the proportional hazards model, joint modeling also incorporates a mixed model to predict an endogenous covariate to enhance the survival model. The joint modeling process, Figure 5-4, involves generating a survival model based on baseline covariates and a longitudinal model to predict a continuous outcome believed to be predictive of the

event. The prediction outcome of the mixed model is then a covariate in the joint model along with the survival model covariates.

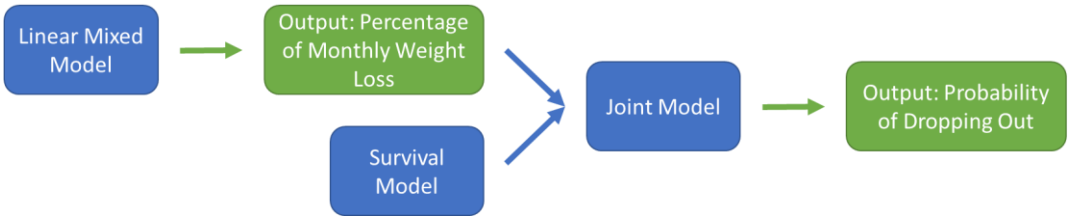

Figure 5-4: Joint model process.

5.3.1 Weight Loss and Drop-Out

To confirm that monthly weight loss percentage is predictive of dropping out of the program, Table 5-7 displays the average monthly weight loss split by members that stayed in the program and those that dropped out of the program in the following month. For example, in Table 5-7, members that are still active in month 2 lost 3.74% of their body weight in month 1. Members that dropped out in month 2 averaged 2.45% weight loss in month 1. Similarly, members that are still active in month 7 lost 0.49% of their body weight in month 6. Members that dropped out in month 7 averaged 0.05% weight gain in month 6. In both examples, a t-test suggests that this is a statistically significant difference between weight loss percentages but does not consider other factors that may influence this relationship. At the $\alpha = 0.05$ level, members at month 8, 9, and 11 do not show a difference in weight loss between those that drop-out and those that stay in the program at months 9, 10, and 12, respectively. This information was drawn from the model development data. Based on these observations there is a larger monthly percentage of weight loss in those members that are staying in the program.

Table 5-7: Average monthly percentage of weight loss by month in the program split by active and non-active members in the next month. Results of separate t-tests conclude significant differences in weight loss between active and non-active members early in the program.

		Monthly Weight Loss Percentage		
Month	N	Active	Dropped Out	P-Value
1	7002	-3.74	-2.45	0.0000
2	6114	-2.87	-1.57	0.0000
3	5315	-1.97	-0.76	0.0000
4	4529	-1.36	-0.58	0.0000
5	3840	-0.83	-0.13	0.0000
6	3278	-0.49	0.05	0.0000
7	2827	-0.25	0.24	0.0000
8	2413	-0.02	0.13	0.1564
9	2082	0.15	0.19	0.7024
10	1818	0.17	0.47	0.0277
11	1594	0.26	0.43	0.1897

5.3.2 JM package

The JM package in R by Dimitris Rizopoulos [48] is utilized to generate the joint model. The JM package requires a *lme* object for the longitudinal model. A *lme* object is created with lme() function within the nlme package [46]. Additionally, the coxph() function in the survival package [49] is used to create the survival model. Utilizing the lme() and coxph() functions in R produces a smooth transition into using the jointModel() function to build the joint model [30]. The jointModel() function extracts the required information from the mixed model and the Cox proportional hazards model to fit the joint model by the maximum likelihood method.

5.3.3 Model

The models described in Section 5.1.3 and Section 5.2.3 were used to fit a joint model. Three joint models are generated to determine the most appropriate model. First, a joint model was generated based on the longitudinal model described in Table 5-4 in Section 5.1.3 and the survival model outlined in Table 5-5 in Section 5.2.3. The joint

model was produced by using a piecewise-constant baseline hazard function with four knots as described in Section 2.2.4.2 in Equation 2-29. The joint model summary is displayed in Table 5-8.

Table 5-8: Joint model output.

Variable	Coefficient	exp(Coefficient)	p-value	Interpretation
Age	-0.0065	0.9935	<0.0001	-0.6%
Marital Status - Relationship	-0.1778	0.8371	0.1154	-16.3%
Marital Status - Single	-0.058	0.9436	0.6086	-5.6%
Gender - Male	0.0211	1.0213	0.5486	2.1%
Assoct	0.8775	2.4049	<0.0001	140.5%
log(xi.1)	-0.4558			
log(xi.2)	-1.0824			
log(xi.3)	-1.5783			
log(xi.4)	-2.1368			
log(xi.5)	11.887			

With a unit increase in age, there is a 0.6% decrease in the risk of dropping out of the program. Members are more likely to stay in the program as age increase. Missing marital status has the highest risk of an event, followed by members that are single, and those that are married are least likely to drop out of Profile. In the survival model, males are described as having 5.7% increase in the risk of dropping out of the weight loss program compared to females. The joint model coefficient changes to only be a 2.1% increase in risk. The sex variable has a p-value of 0.55 in the joint model (sex covariate p-value in survival model is 0.04) which is no longer considered significant in the model now that the longitudinal outcome of monthly weight loss percentage has been introduced into the model. There are similarities and differences in the coefficients of the joint model in Table 5-8 to those in Table 5-5 for the survival model. A direct comparison is shown in Table 5-9. All parameter estimates have changed slightly but all coefficient signs remain the same.

Table 5-9: Comparison of survival and joint model coefficients.

Variable	Coefficients	
	Survival	Joint
Age	-0.0135	-0.0065
Marital Status - Relationship	-0.2498	-0.1778
Marital Status - Single	-0.0828	-0.0580
Sex - Male	0.0553	0.0211

Added to the joint model, Table 5-8, is the *Assoct* parameter and several parameters for the piecewise baseline hazards function. The *Assoct* parameter refers to α in Equation 2-24. This parameter describes the relationship between the longitudinal outcome and the risk of an event. With a unit increase in the percentage of monthly weight loss (weight gain), the risk of dropping out increases by 140%, with everything else held constant.

The JM package and jointModel() function allow for extensions of the joint model as described in Section 2.2.4.2. Both extensions were applied to the joint model and the resulting AIC, AUC and accuracy are shown in Figure 5-5 shows a graphical representation of the Receiver Operating Characteristic (ROC) curves for all three models as applied to the validation data.

Table 5-10: Comparison of three joint models.

Model	AIC	Validation		
		AUC	Accuracy	True Positive
Joint Model	176507	0.6320	80.66	2.48
Slope	176235	0.6324	81.00	2.44
Cumulative	176170	0.6254	80.79	2.42

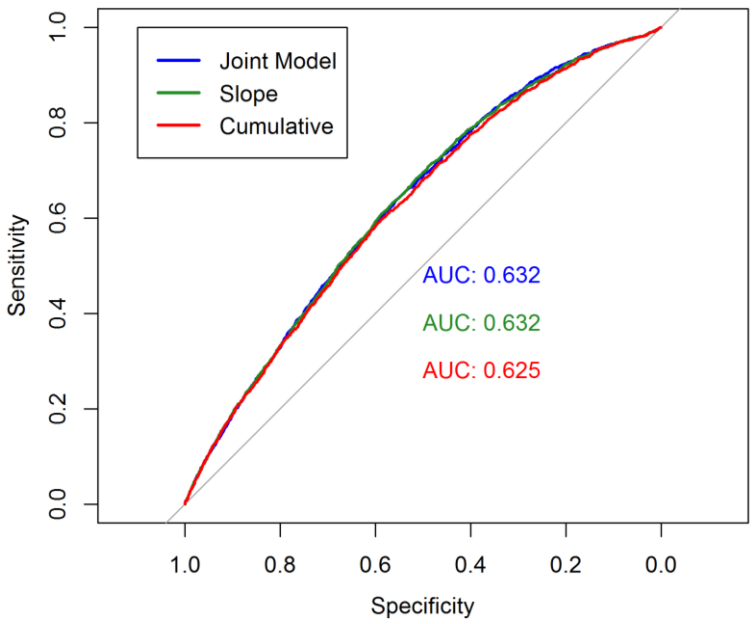

Figure 5-5: Comparison of ROC curves and AUC values for three joint models.

The two extensions are expected to improve the model performance. As shown in Table 5-10, the cumulative model has the best AIC, the slope model has the highest AUC and accuracy measurements, and the original joint model has the highest true positive value. All accuracy and true positive measurements are based on cutoff points for the classification of survival probabilities. The separation between each of these measurements is small. In Figure 5-5, all three ROC curves are nearly on top of each other with the cumulative model being slightly lower than the other two. With these considerations, the original joint model without any extensions will be the final model. The coefficients of this model were displayed in Table 5-8 and discussed earlier.

Table 5-11: Comparison of the survival model and the joint model on the validation data.

Model	Validation		
	AUC	Accuracy	True Positive
Survival Model	0.5616	78.9	2.37
Joint Model	0.6320	80.7	2.48

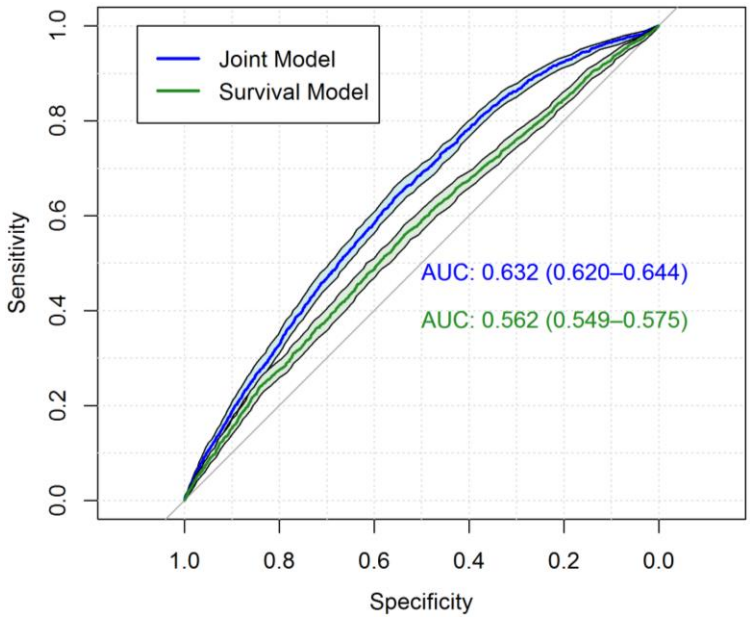

Figure 5-6: Comparison of ROC curves and AUC values for the survival and joint models. Included is the 95% confidence interval around each ROC curve and the 95% confidence interval is printed next to each AUC value.

Since the joint model is an enhancement of the survival model, we can compare the two to determine if, indeed, an improvement was made. Table 5-11 and Figure 5-6 show a direct comparison of the survival model and joint model. Figure 5-6 shows the two ROC curves with AUC values and their corresponding 95% confidence intervals. All measurements in Table 5-11 are larger in the joint model compared to the survival model. It is also apparent by the ROC curves that the joint model is better at classifying members by their survival probabilities than the survival model.

5.4 Application of the Joint Model

Shiny is an R package that allows users to create interactive web applications [50]. The package has access to prebuilt widgets which can input values and output results to make a professional looking application with minimal effort [51]. Applications are customizable to display tables and graphs however the user chooses.

Shiny is used to create an application that can be utilized by Profile coaches. The application uses the models described throughout Chapter 5. Currently, it is built to handle members that are actively involved in the weight loss program. The model is not built to handle situations in which a member drops out of the program but continues in a subsequent month. The application inputs the member's identification number and the output includes a four-tab panel.

The first tab, labeled *Graphs*, includes a text box to input the member's identification number. The layout of this tab is shown in Figure 5-7. After submitting the member's identification number, two graphs appear. The top graph displays the member's weight from month 1 to month 12. If the line is blue, this indicates an actual measured weight and if the line is red, this indicates the projected weight. The mixed model described in Section 5.1.3 is used to generate a projected monthly percentage of weight loss and this projection is converted to an actual weight in pounds. The bottom graph shows the probability of dropping out of the program over time. The blue line will always be at 0 since this indicates actual value and since the member is still active, they have not dropped out. The red line indicates the probability of dropping out the program based on the joint model described in Section 5.3.3. Profile coaches would have access to this information to determine the likelihood of the member staying in the program.

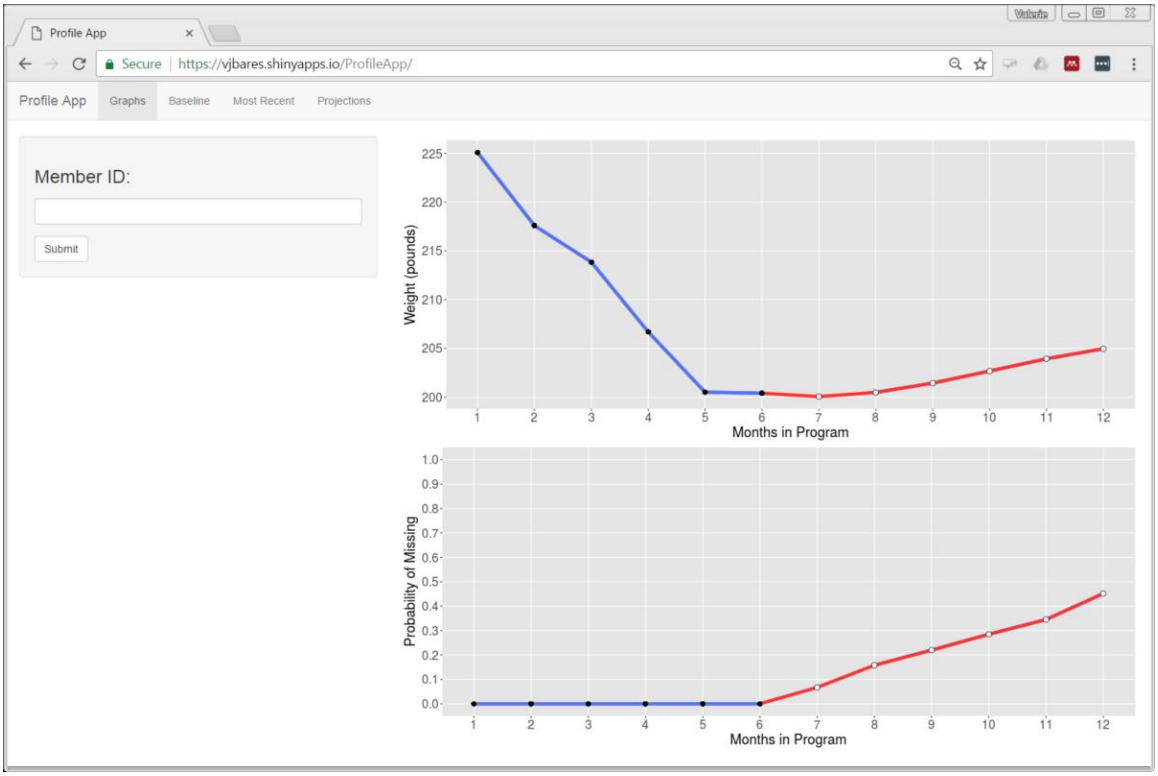

Figure 5-7: *Graphs* tab of Shiny app.

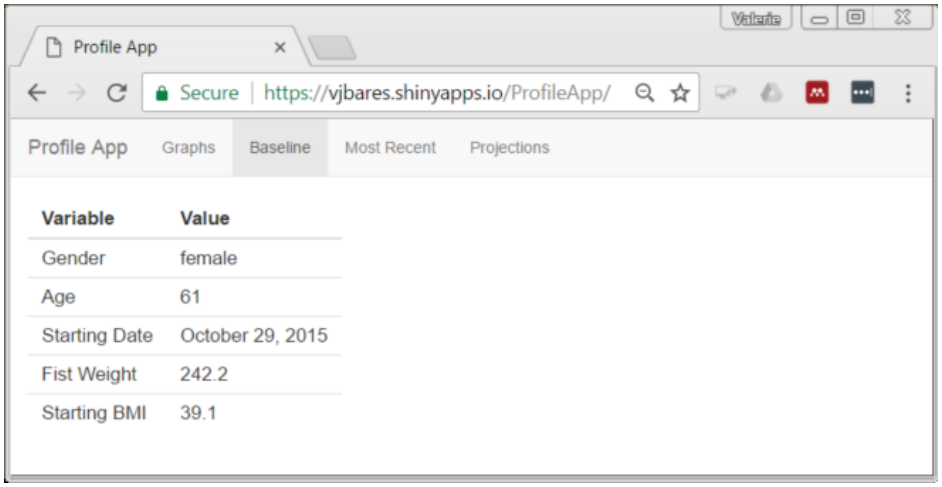

Figure 5-8: *Baseline* tab of Shiny app.

Figure 5-8 shows the second tab in the application, *Baseline*. Information from when the member started Profile including their sex, age, starting date, starting weight, and starting BMI is displayed. The third tab, *Most Recent*, is shown in Figure 5-9 and

displays the last known information for the member. This tab can inform the Profile coach with pertinent information about the member including their current weight, current BMI, how many coach meetings they attended by the end of the previous month, how many weight recordings they had by the end of the last month, and information about how much weight they have currently lost. This information gives the coach a quick look at the member's behavior and their progress.

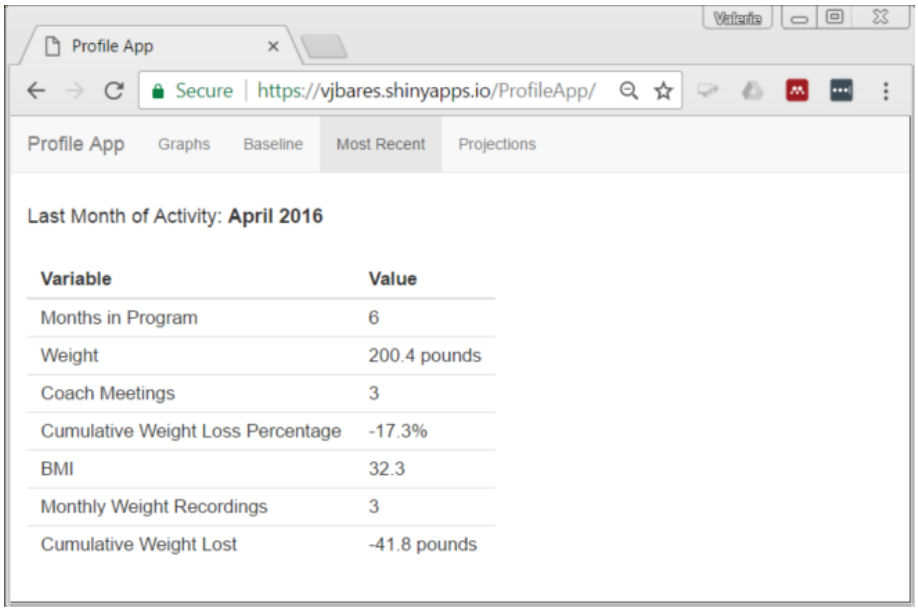

Figure 5-9: *Most Recent* tab of Shiny app.

The last tab, *Projections*, is used to look at the member's past behavior or projected behavior. Figure 5-10 and Figure 5-11 display the same member with a different value chosen in the drop-down box. The drop-down box at the top of the page allows the coach to choose which month they want to examine. For example, Figure 5-10 displays information for the member 6 months into the program. Text at the top of the page indicates "These values are *actual* values." as opposed to projected values. The

coach can examine the percentage of weight loss the member had in month 6 (monthly and cumulative) as well as their weight and BMI.

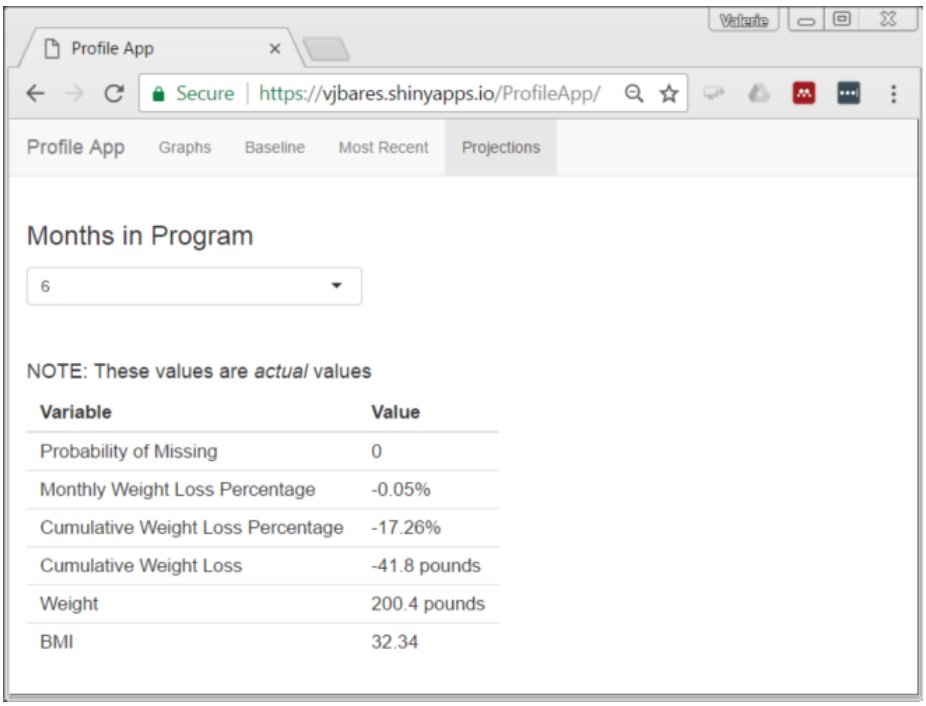

Figure 5-10: *Projections* tab of Shiny app with actual values.

The coach may also be interested in the projections for month 12 for the same member. By clicking on the drop-down box, the number 12 can be selected and the information changes accordingly, as shown in Figure 5-11. The text at the top of the page indicates "These values are *projected* values." as opposed to actual values. This member has a 0.4516 probability of dropping out of the program by month 12 and projected to lose a cumulative 15.38% of their weight. This information is analogous to the graphs in the first tab.

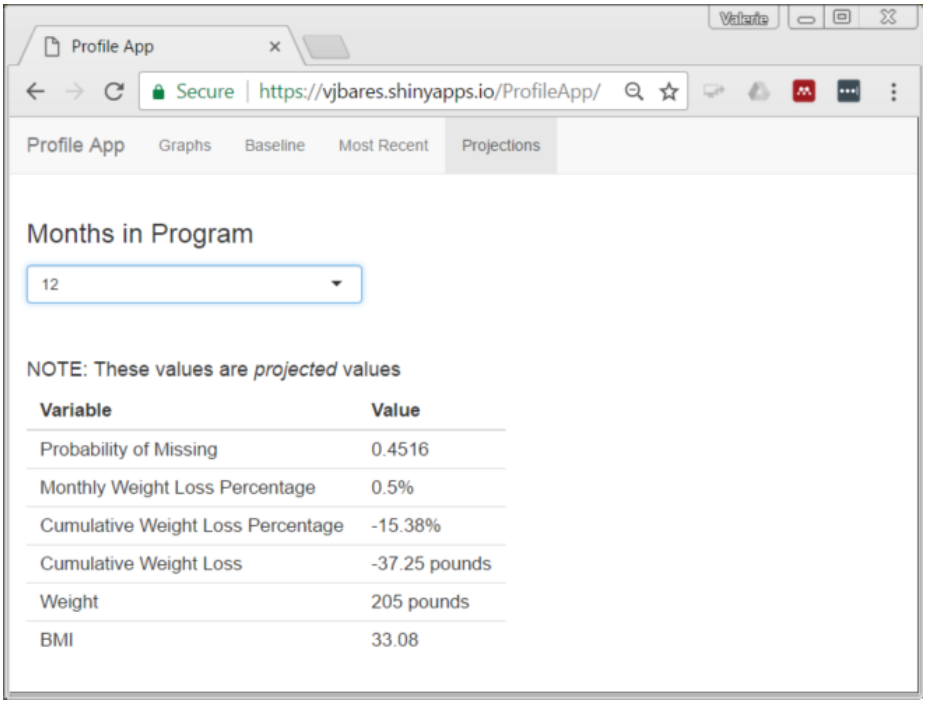

Figure 5-11: *Projections* tab of Shiny app with projected values.

5.5 CONCLUSIONS

Predicting a member's body weight based on their behavior indicates that by attending meetings with a weight loss coach and keeping track of your weight is predictive of more weight loss. With each additional coach meeting a member loses 0.13 percentage points more monthly and with one more weight recording they lose 0.02 percentage points more per month. A unit increase in starting BMI is associated with an increase of 0.03 percentage points more weight loss.

By month 6 less than half of members are still in the program and by month 12 only 20% are still in the program. A relationship is observed between weight loss and members dropping out of Profile over time. Members that are still active in month 7 lost 0.49% of their body weight in month 6 whereas members that dropped out in month 7 averaged 0.05% weight gain in month 6.

Joint models generated to predict the probability a member will drop out of Profile shows that a unit increase in age is associated with a 0.6% decrease in the risk of dropping out. Missing marital status has the highest risk of dropping out, followed by members that are single and then those that are married are the least likely to drop out of Profile. Males are associated with a 2.1% higher risk of dropping out of Profile compared to females. The joint model associates a 140% increased risk of dropping out with each percentage point increase in monthly weight gain. The likelihood of a member dropping out of the program increases with less weight loss.

6 SUMMARY

6.1 DISCUSSION

Profile by Sanford data was utilized to discover trends in member behavior. Available data included basic demographic information, weight measurements, body part measurements, coach interactions, recorded exercise and food consumption, meal plan information, and medication use. Data was collected in various ways such as coach or member input and body weight scale recordings.

Several statistical methods are applied to the Profile data. Exploratory data analysis is used to discover variable trends over time and relationships between variables. Linear regression is utilized to examine the association of coach meeting frequency to weight loss at month 12. A mixed effects model predicts weight loss over time by selected covariates. A semiparametric survival model utilizes covariates to predict the probability of dropping out of the Profile program. Joint models are also used to predict the probability of dropping out of the program but this method combines the mixed model and the survival model. Joint models enhance the Cox proportional hazards model by including the prediction from the mixed model into the survival model.

Not all Profile members utilize tools they are provided such as recording exercise and food consumption. An increase in the use of these tools could add additional understanding of the relationship of member behavior and weight loss. Organized data collection on coach interactions would support additional associations. Indication whether a meeting was face-to-face or virtual could give further insight into these interactions. Joint modeling to associate a time-dependent covariate to a risk of an event can be computationally extensive. Computation time grows with increasingly complex

mixed and survival models. This limitation can prevent the creation of robust mixed and survival models which suppresses the predictive power of the joint model.

More detailed information is linked to Profile which could enhance both the mixed model and the survival model. Purchase information, including whether the member has purchased food items through Profile, may improve the mixed model. The joint model could also be created to project weekly weight loss and drop-out probabilities. The current joint model only handles members that have not missed a weight measurement (not dropped-out). Enhancements to the model can be made to include these members and include this information as an indicator of future behavior.

The web application could be enhanced by allowing sliders or additional drop-down boxes to change inputs of the mixed model. This could give insight into how and when a member could reach their weight loss goal. Applications of the joint model could include proactive retention strategies. Automated mailing or e-mailing strategies could be built to motivate members that have a high probability of dropping out of the program.

6.2 CONCLUSIONS

Members of Profile by Sanford are 77% female and two-thirds start the program as obese. These members live in 48 of the 50 states. Only 20% of members have utilized Profile to record their exercise and 57% have logged food items. Members record their weight, on average, 7 to 8 times a month and 68% of those measurements are between 5 AM and 9 AM. After 12 months in the program, members that have recorded a weight have lost an average of 10.8% of their body weight.

Increasing a member's coach meeting attendance to one more meeting a month results in 2.5 percentage points more weight loss for Profile members who weigh themselves consistently each month for the first 12 months in the program. The same group of Profile members are seeing 2.3 percentage points less weight loss if taking antidepressants after controlling for sex and starting BMI. For this subset of members, while holding antidepressant use and starting BMI constant, females are losing 1.9 percentage points more than males.

A higher frequency of monthly coach meetings, more monthly weight recordings, and a higher starting BMI are all predictive of greater monthly weight loss percentage. By month 6 less than half of members are still in the program and by month 12 only 20% are still in the program. Higher age, married members, and females are associated with a lower risk of dropping out of Profile. The joint model associates a 140% increased risk of dropping out with each percentage point increase in monthly weight gain. The likelihood of a member dropping out of the program increases with less weight loss. The area under the ROC curve measures the ability of a model to classify members by their predicted probabilities. The area under the ROC curve for the Cox proportional hazards model is 0.5616 and the joint model is 0.6320. When directly comparing the predictability of the survival and joint model, it is evident that the joint model classifies members by their survival probabilities better than the survival model.

APPENDIX

Github repository for R code: https://github.com/vjbares/Profile

REFERENCES

1. Adult Obesity Facts | Overweight & Obesity | CDC [Internet]. [cited 2016 May 24]. Available from: http://www.cdc.gov/obesity/data/adult.html

2. Gudzune KA, Doshi RS, Mehta AK, Chaudhry ZW, Jacobs DK, Vakil RM, et al. Efficacy of Commercial Weight-Loss Programs: An Updated Systematic Review. Ann. Intern. Med. 2015;162:501–12.

3. Ogden CL, Carroll MD, Kit BK, Flegal KM, LK K, CL O, et al. Prevalence of Childhood and Adult Obesity in the United States, 2011-2012. JAMA. American Medical Association; 2014;311:806–14.

4. Body mass index - BMI [Internet]. World Health Organization; [cited 2017 Jun 13]. Available from: http://www.euro.who.int/en/health-topics/disease-prevention/nutrition/a-healthy-lifestyle/body-mass-index-bmi

5. Body Mass Index (BMI) [Internet]. Centers Dis. Control Prev. 2015 [cited 2017 Jun 13]. Available from: https://www.cdc.gov/healthyweight/assessing/bmi/

6. Müller MJ, Braun W, Enderle J, Bosy-Westphal A. Beyond BMI: Conceptual Issues Related to Overweight and Obese Patients. Obes. Facts. 2016;9:193–205.

7. Assessing Your Weight and Health Risk [Internet]. Natl. Hear. Lung, Blood Inst. Available from: https://www.nhlbi.nih.gov/health/educational/lose_wt/risk.htm#limitations

8. World Health Organization. Waist Circumference and Waist-Hip Ratio: Report of a WHO Expert Consultation. 2011.

9. Finkelstein EA, Trogdon JG, Cohen JW, Dietz W. Annual Medical Spending Attributable To Obesity: Payer-And Service-Specific Estimates. Health Aff. 2009;28:822–31.

10. Sacks FM, Bray GA, Carey VJ, Smith SR, Ryan DH, Anton SD, et al. Comparison of Weight loss Diets with Different Compositions of Fat, Protein and Carbohydrates. N. Engl. J. Med. 2009;360:859–73.

11. Clifton PM, Condo D, Keogh JB. Long term weight maintenance after advice to consume low carbohydrate, higher protein diets - A systematic review and meta analysis. Nutr. Metab. Cardiovasc. Dis. Elsevier Ltd; 2014;24:224–35.

12. Soenen S, Bonomi AG, Lemmens SGT, Scholte J, Thijssen M a M a, Van Berkum F, et al. Relatively high-protein or "low-carb" energy-restricted diets for body weight loss and body weight maintenance? Physiol. Behav. 2012;107:374–80.

13. Anastasiou C a., Karfopoulou E, Yannakoulia M. Weight regaining: From statistics and behaviors to physiology and metabolism. Metabolism. Elsevier Inc.; 2015;64:1395–407.

14. Wing RR, Tate DF, Gorin A a, Raynor H a, Fava JL. A Self-Regulation Program for Maintenance of Weight Loss. N. Engl. J. Med. 2006;355:1563–71.

15. Holzapfel C, Merl M, Stecher L, Hauner H. One-Year Weight Loss with a Telephone-Based Lifestyle Program. Obes. Facts. 2016;9:230–40.

16. Sherwood NE, Jeffery RW, Welsh EM, Vanwormer J, Hotop AM. The Drop It At Last (DIAL) Study: Six month results of a phone-based weight loss trial. Am. J. Heal. Promot. NIH Public Access; 2010;24:378–83.

17. MacLean PS, Wing RR, Davidson T, Epstein L, Goodpaster B, Hall KD, et al. NIH Working Group Report: Innovative Research to Improve Maintenance of Weight Loss. Obesity. 2015;23:7–15.

18. Sciamanna CN, Kiernan M, Rolls BJ, Boan J, Stuckey H, Kephart D, et al. Practices Associated with Weight Loss Versus Weight-Loss Maintenance. Am. J. Prev. Med. 2011;41:159–66.

19. Ramage S, Farmer A, Eccles KA, Mccargar L. Healthy strategies for successful weight loss and weight maintenance : a systematic review. Appl. Physiol. Nutr. Metab. - NRC Res. Press. 2014;39:1–20.

20. Thomas DM, Ivanescu AE, Martin CK, Heymsfield SB, Marshall K, Bodrato VE, et al. Predicting successful long-term weight loss from short-term weight-loss outcomes: new insights from a dynamic energy balance model (the POUNDS Lost study). Am. J. Clin. Nutr. 2015;101:449–54.

21. Ortner Hadziabdic M, Mucalo I, Hrabac P, Matic T, Rahelic D, Bozikov V. Factors predictive of drop-out and weight loss success in weight management of obese patients. J. Hum. Nutr. Diet. 2015;28:24–32.

22. Sawamoto R, Nozaki T, Furukawa T, Tanahashi T, Morita C, Hata T, et al. Predictors of Dropout by Female Obese Patients Treated with a Group Cognitive Behavioral Therapy to Promote Weight Loss. Obes. Facts. 2016;9:29–38.

23. Profile by Sanford [Internet]. [cited 2017 Jun 20]. Available from: https://www.profileplan.net/

24. R Core Team. R: A language and environment for statistical computing [Internet]. Vienna, Austria: R Foundation for Statistical Computing; 2017. Available from: https://www.r-project.org/

25. Wickham H. ggplot2: Elegant Graphics for Data Analysis [Internet]. New York: Springer-Verlag; 2009. Available from: http://ggplot2.org

26. Ripley B, Lapsley M. RODBC: ODBC Database Access [Internet]. 2017. Available from: https://cran.r-project.org/package=RODBC

27. Lyche T, Morken K. Spline Methods [Internet]. University of Oslo, Department of Informatics, Centre of Mathematics for Applications; 2008. Available from: http://www.uio.no/studier/emner/matnat/ifi/INF-MAT5340/v10/undervisningsmateriale/book.pdf

28. de Boor C. B(asic)-Spline Basics [Internet]. Available from: https://www.cs.unc.edu/~dm/UNC/COMP258/Papers/bsplbasic.pdf

29. Crowther MJ, Abrams R, Lambert PC. Flexible parametric joint modelling of longitudinal and survival data. Stat. Med. 2012;31:4456–4471.

30. Rizopoulos D. Joint Models for Longitudinal and Time-to-Event Data: With Applications in R. Boca Raton: Chapman & Hall/CRC; 2012.

31. Stroup WW. Generalized Linear Mixed Models: Modern Concepts, Methods and Applications. Boca Raton: Chapman & Hall/CRC; 2013.

32. Moore DF. Applied Survival Analysis Using R. Springer; 2016.

33. Fox J, Weisberg S. Cox Proportional-Hazards Regression for Survival Data in R. An Append. to An R Companion to Appl. Regression, Second Ed. 2011;1–20.

34. Wulfsohn MS, Tsiatis AA. A Joint Model for Survival and Longitudinal Data Measured with Error. Biometrics. 1997;53:330–9.

35. Elashoff RM, Li G, Li N. Joint Modeling of Longitudinal and Time-to-Event Data. Boca Raton: Chapman & Hall/CRC; 2017.

36. Rizopoulos D. Dynamic Predictions and Prospective Accuracy in Joint Models for Longitudinal and Time-to-Event Data. Biometrics. 2011;67:819–29.

37. Hickey GL, Philipson P, Jorgensen A, Kolamunnage-dona R. Joint modelling of time-to-event and multivariate longitudinal outcomes: recent developments and issues. BMC Med. Res. Methodol. 2016;16:117.

38. Rizopoulos D. JM : An R Package for the Joint Modelling of Longitudinal and Time-to-Event Data. J. Stat. Softw. 2010;35.

39. Kantor ED, Rehm CD, Haas JS, Chan AT, Giovannucci EL. Trends in Prescription Drug Use Among Adults in the United States From 1999-2012. JAMA. 2015;314:1818–31.

40. National Center for Health Statistics. Health, United States, 2015: With Special Feature on Racial and Ethnic Health Disparities [Internet]. Hyattsville, MD; 2016. Available from: http://www.ncbi.nlm.nih.gov/pubmed/27308685

41. VanderWeele TJ. The Sign of the Bias of Unmeasured Confounding. Biometrics. 2008;64:702–6.

42. VanderWeele TJ, Shpitser I. On the Definition of a Confounder. Ann. Stat. 2013;41:196–220.

43. Martin LJ. High blood pressure medicines [Internet]. U.S. Natl. Libr. Med. 2015 [cited 2017 Jun 1]. Available from: https://medlineplus.gov/ency/article/007484.htm

44. Types of Blood Pressure Medications [Internet]. Am. Hear. Assoc. 2016 [cited 2017 Jun 1]. Available from: http://www.heart.org/HEARTORG/Conditions/HighBloodPressure/MakeChangesThatMatter/Types-of-Blood-Pressure-Medications_UCM_303247_Article.jsp#.WTBzZevytb4

45. Sara R, Pai NB, Deng C. The Association of Antidepressant Medication and Body Weight Gain. Online J. Heal. Allied Sci. [Internet]. 2013;12:1–9. Available from: http://www.ojhas.org/issue45/2013-1-1.html

46. Pinheiro J, Bates D, DebRoy S, Sarkar D. nlme: Linear and Nonlinear Mixed Effects Models [Internet]. 2017. Available from: https://cran.r-project.org/package=nlme

47. Singh R, Mukhopadhyay K. Survival analysis in clinical trials: Basics and must know areas. Perspect. Clin. Res. Medknow Publications; 2011;2:145–8.

48. Rizopoulos D. JM: An R Package for the Joint Modelling of Longitudinal and Time-to-Event Data. J. Stat. Softw. [Internet]. 2010;35:1–33. Available from: http://www.jstatsoft.org/v35/i09/

49. Therneau TM. A Package for Survival Analysis in S [Internet]. 2015. Available from: https://cran.r-project.org/package=survival

50. Shiny by RStudio: A web application framework for R [Internet]. 2016 [cited 2017 Jun 5]. Available from: https://shiny.rstudio.com/

51. Chang W, Cheng J, Allaire J, Xie Y, McPherson J. shiny: Web Application Framework for R [Internet]. 2017. Available from: https://cran.r-project.org/package=shiny